

Model surgery: joining and splitting models with Markov melding

Robert J. B. Goudie, Anne M. Presanis, David Lunn,
Daniela De Angelis and Lorenz Wernisch

September 5, 2016

MRC Biostatistics Unit, Cambridge, UK

Abstract

Analysing multiple evidence sources is often feasible only via a modular approach, with separate submodels specified for smaller components of the available evidence. Here we introduce a generic framework that enables fully Bayesian analysis in this setting. This requires *joining* the submodels together into a full, joint model, so that all uncertainty can be fully propagated between all components. We propose a generic method for forming a suitable joint model from submodels, and a convenient computational algorithm for fitting this joint model in stages, rather than as a single, monolithic model. The approach also enables *splitting* of large joint models into smaller submodels, allowing inference for the original joint model to be conducted via our multi-stage algorithm. We motivate and demonstrate our approach through two examples: joining components of an evidence synthesis of A/H1N1 influenza, and splitting a large ecology model.

KEYWORDS: Markov combination; Bayesian melding; evidence synthesis; model integration

1 Introduction

The synthesis of evidence from multiple sources and from different study designs is increasingly common in all areas of science, including ecology (Clark *et al.*, 2010), biomedical kinetics (Henderson *et al.*, 2010), infectious disease epidemiology (Birrell *et al.*, 2011) and health technology assessment (Welton *et al.*, 2012). However, dealing with joint models of several sources of evidence, including data and expert opinion, may be inferentially and computationally challenging, or even infeasible. It is often sensible, if not necessary, to take a modular approach by considering separate submodels for smaller, more manageable, components of the available evidence (Green *et al.*, 2003; Liu *et al.*, 2009). These submodels can originate in two ways: either by first specifying submodels that, in a Bayesian framework, should be *joined* into a single model to allow all information and uncertainty to be fully propagated; or as a result of *splitting* an existing joint model. Key advantages of a modular approach are computational efficiency and facilitation of model criticism, since submodels might provide conflicting inference on common parameters (Presanis *et al.*, 2013; Gåsemeyr and Natvig, 2009).

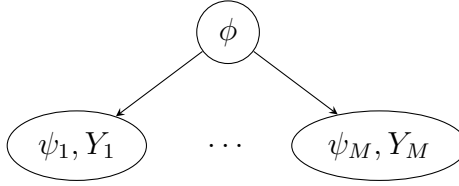


Figure 1: DAG representation of a joint hierarchical model linking M submodels.

Formally, consider M probability submodels $p_m(\phi, \psi_m, Y_m)$, $m = 1, \dots, M$, for submodel-specific multivariate random variables ψ_m and Y_m , as well as a multivariate random variable ϕ common to all modules that acts as a ‘link’ between the submodels. We would like to *join* the submodels into a single model $p(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M)$ so that the posterior distributions for the link parameter ϕ and the submodel-specific parameters ψ_m account for all observations and uncertainty. A suitable joint model for a collection of submodels naturally arises in some contexts from standard model constructs, such as a hierarchical model (Figure 1). However, it is not immediately clear how to form such a joint model when either: the submodels are not expressed in a form conditional upon the link parameter ϕ ; the linking parameter is deterministically related to other parameters in one of the submodels; or the marginal distributions $p_m(\phi)$, $m = 1, \dots, M$, for the link parameter ϕ differ in the submodels. In applied research, convenient approximate two-stage approaches have been widely used, where one submodel is fitted and an approximation of the resulting posterior is provided to a second submodel (Jackson *et al.*, 2009; Presanis *et al.*, 2014). However, the joint model that is implied by such an approach is unclear (Eddy *et al.*, 1992; Ades and Sutton, 2006).

Conversely, suppose we have an existing joint model $p(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M)$ that we wish to *split* into M submodels $p_m(\phi, \psi_m, Y_m)$, $m = 1, \dots, M$. The submodels should be faithful to the original model in the sense that joining the submodels results in the original model. In some contexts, suitable submodels arise naturally from the structure of the joint model, resulting in splitting strategies used implicitly in the context of hierarchical models (Lunn *et al.*, 2013a; Tom *et al.*, 2010; Liang and Weiss, 2007) and of tall data (Scott *et al.*, 2016; Neiswanger *et al.*, 2014). However, neither the general conditions stipulating when splitting is permissible nor a general framework for splitting a model are immediately clear.

In this paper we introduce *Markov melding*, a simple, generic approach for joining and splitting models that brings together, rephrases, clarifies and generalises various proposals made in the literature under the umbrella of one theoretical framework. In terms of joining, Markov melding extends the theoretical framework of *Markov combination* (Dawid and Lauritzen, 1993) with ideas from *Bayesian melding* (Poole and Raftery, 2000), enabling evidence synthesis (Ades and Sutton, 2006; Welton *et al.*, 2012) and model expansion (Draper, 1995; Gelman *et al.*, 2014) in realistic applied settings. Specifically, while a prerequisite for Markov combination is the restrictive condition of identical prior marginal distributions $p_m(\phi)$, $m = 1, \dots, M$, our approach enables joining of submodels with similar but not identical marginal distributions. Markov melding aims to preserve the original submodels as faithfully as possible, and, in particular, always preserves the submodel-specific conditional distributions $p_m(\psi_m, Y_m \mid \phi)$ for all $m = 1, \dots, M$. We also account for contexts where a deterministic relationship between the link parameter ϕ and other parameters is not invertible, allowing therefore for model joining in situations that have traditionally been challenging. In terms of splitting, our Markov melding

framework clarifies the conditions required and the general framework in which to conduct model splitting, facilitating the modular approach advocated above. Notably, we generalise existing tall data splitting approaches (Scott *et al.*, 2016; Neiswanger *et al.*, 2014) for independent, identically distributed data to other types of data.

Finally, we also develop an algorithm for fitting the Markov melded model in stages, rather than as a single monolithic model, for both joining and splitting models. The algorithm proposed here extends naturally that employed in Lunn *et al.* (2013a) and is closely related to those used in Liang and Weiss (2007) and Tom *et al.* (2010).

The paper is organised as follows: in Section 2 we introduce some examples motivating this work; Section 3 provides the conceptual framework underlying our approach; inferential and computational aspects of the approach are presented in Section 4; Section 5 gives details and results for the motivating examples; we conclude with a discussion and suggestions for further work in Section 6.

2 Motivating examples

We motivate and demonstrate our framework for joining and splitting models with two examples, providing here a brief high-level outline of each. In Section 3.5 we will describe how Markov melding applies in each case. We defer full details of each model until Section 5.

For both examples, as in the rest of the paper, we use directed acyclic graphs (DAGs) to represent the dependence structure between variables in a model (Figures 2 and 3). Each variable in the model is represented by a node with rectangular nodes denoting observed variables and links between the nodes indicating direct dependencies. Stochastic dependencies are represented by solid lines and deterministic (logical) relationships by dashed lines. The joint distribution of all nodes is the product of the conditional distributions of each node given its direct parents, and conditional independence relationships can be read from the graph (Lauritzen, 1996).

2.1 Joining: A/H1N1 influenza evidence synthesis

Any public health response to influenza outbreaks relies on knowledge of severity: the probability that an infection results in a severe event such as hospitalisation or death. Relevant data, however, are often sparse, biased and originally collected for other purposes. Estimation therefore requires the coherent synthesis of a range of complementary but disparate sources of information (De Angelis *et al.*, 2015). A high-level representation of a typical, although simplified, severity model is shown in Figure 2, based on a synthesis of evidence relating to the 2010-2011 A/H1N1 influenza outbreak in the UK (Presanis *et al.*, 2014). We consider evidence from an intensive care unit (ICU) submodel and how it is combined with the rest of the evidence synthesis to estimate severity.

The ICU submodel is an immigration-death process, governed by parameters θ representing transition rates in and out of ICU. The submodel combines weekly counts y of *suspected* influenza cases with information on the proportion π^{pos} of tested cases that are positive for the A/H1N1 strain, to estimate the latent weekly incident new admissions to ICU of A/H1N1 cases. The severity submodel synthesises several datasets, each of which is only available aggregated over time, to estimate the *cumulative* incident ICU admissions of A/H1N1, χ . The different timescales in the two submodels necessitate summing the ICU submodel weekly A/H1N1 admissions over time, to obtain an estimate of the cumu-

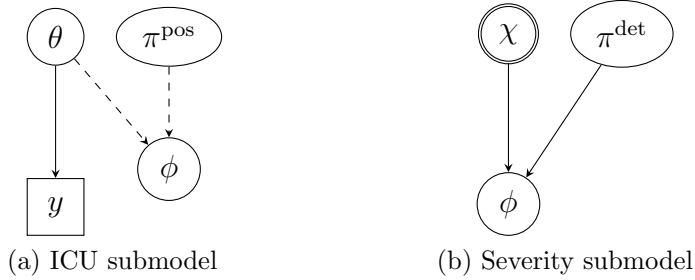


Figure 2: High-level DAG representations of the influenza submodels. The double circle denotes the (highly) informative prior for χ , reflecting data from the full severity submodel that is omitted here. Detailed DAGs of these submodels are shown in Figure 6.

lative admissions ϕ . ϕ is therefore a deterministic function $\phi(\theta, \pi^{\text{pos}})$ of the ICU submodel parameters. A further complication is that the severity submodel datasets cover a longer time period than the ICU data, so ϕ is an under-estimate of χ . A binomial model therefore relates the two quantities, accounting for the difference via the probability parameter π^{det} . The full severity submodel is more complex than shown, appearing originally as parents of χ in Figure 2b. Here, this complexity is, for simplicity, summarised only through an informative prior for χ .

A joint model that accounts for the different timescales in the two submodels is challenging to formulate, since the deterministic function $\phi(\theta, \pi^{\text{pos}})$ is not invertible. Presanis *et al.* (2014) therefore transferred information between these two separate submodels via an approximate approach (see Section 4.3 for details). We show in this paper how Markov melding can be used to join the two submodels formally into a single joint model, making all the assumptions involved explicit. We also explain the relationship between our approach and the standard approximate approach.

2.2 Splitting: large ecology model

As an example of splitting a large DAG model we consider a joint model (Besbeas *et al.*, 2002) for two distinct sources of data about British Lapwings (*Vanellus vanellus*). These data sources are primarily collected to inform different aspects of studies of the birds: census-type data provide a measure of breeding population size, while mark-recapture-recovery data provide estimates of the annual survival probability of the birds via observations of the survival of uniquely marked individuals. These data are related and a joint model allows inference to account for all information available. In the joint Bayesian model of Brooks *et al.* (2004) (Figure 3a), the mark-recapture-recovery data y are modelled in terms of the recovery rate λ , and the survival rate ϕ for birds; and the census data x are modelled in terms of the survival rate ϕ , and the productivity rate γ of adult female birds. The joint model links the data sources using the common survival rate parameter (ϕ).

Brooks *et al.* (2004) considered fitting the census and mark-recapture-recovery models both separately and jointly using standard Markov chain Monte Carlo (MCMC) algorithms. We describe in this paper how, through Markov melding, inference from such a joint model can be carried out in stages after splitting the model into two separate components (Figure 3b), circumventing the need to directly fit the full, joint model in a single MCMC procedure.

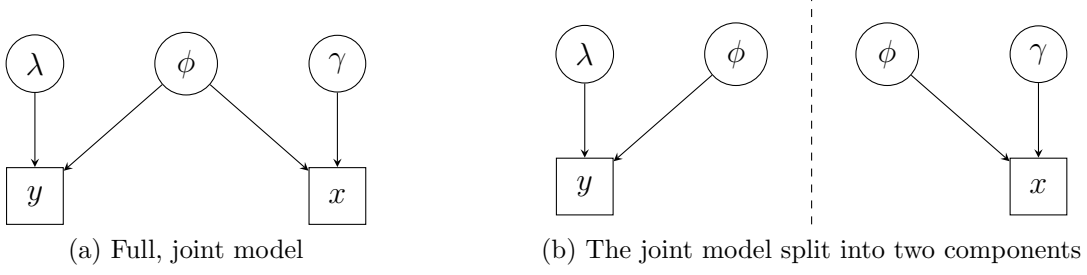


Figure 3: High-level DAG representations of the ecology models. Detailed DAG representations of these models are shown in Figure 9.

3 Conceptual framework

3.1 Notation and assumptions

Let p denote either a probability distribution for discrete random variables or a probability density for continuous variables (we assume such a density exists). In both cases we talk of p interchangeably as a probability or probability distribution and we express conditional probabilities as $p_m(\psi_m | \phi) = p_m(\psi_m, \phi) / p_m(\phi)$, where $p_m(\phi) > 0$. For random variables X_1 , X_2 and X_3 we write $X_1 \perp\!\!\!\perp X_2 | X_3$ when X_1 and X_2 are conditionally independent given X_3 . In general we denote a vector with components x_i , $i = 1, \dots, n$ by $(x_i)_{i=1}^n$.

3.2 Markov combination

To start with, consider the $M = 2$ case. Our aim is to define a joint probability model $p_{\text{comb}}(\phi, \psi_1, \psi_2, Y_1, Y_2)$ that ‘joins’ together two submodels $p_1(\phi, \psi_1, Y_1)$ and $p_2(\phi, \psi_2, Y_2)$ sharing a common variable ϕ . We assume that ϕ has the same marginal distribution in each submodel $p_1(\phi) = p_2(\phi) = p(\phi)$. A sensible requirement on the joint model is that the marginal distributions of ϕ , ψ_m and Y_m in the joint model agree with the marginals in the original submodels, that is, $p_{\text{comb}}(\phi, \psi_m, Y_m) = p_m(\phi, \psi_m, Y_m)$, $m = 1, 2$. The contribution of ϕ to each submodel can be obtained by conditioning: $p_m(\phi, \psi_m, Y_m) = p_m(\psi_m, Y_m | \phi) p_m(\phi)$. This suggests a simple way to join the two submodels while preserving marginals:

$$p_{\text{comb}}(\phi, \psi_1, \psi_2, Y_1, Y_2) = p(\phi) p_1(\psi_1, Y_1 | \phi) p_2(\psi_2, Y_2 | \phi)$$

Marginalising over (ψ_2, Y_2) shows that $p_{\text{comb}}(\phi, \psi_1, Y_1) = p_1(\phi, \psi_1, Y_1)$ and similarly for $p_{\text{comb}}(\phi, \psi_2, Y_2)$.

Markov combination (Dawid and Lauritzen, 1993) generalises this approach to $M \geq 2$. A collection of submodels $p_m(\phi, \psi_m, Y_m)$, $m = 1, \dots, M$, are *consistent* in the link parameter ϕ if the marginal distribution $p_m(\phi) = p(\phi)$ is the same for all $m = 1, \dots, M$. The *Markov combination* p_{comb} of M consistent submodels is then defined as the following joint model:

$$\begin{aligned} p_{\text{comb}}(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M) &= p(\phi) \prod_{m=1}^M p_m(\psi_m, Y_m | \phi) \\ &= \frac{\prod_{m=1}^M p_m(\phi, \psi_m, Y_m)}{p(\phi)^{M-1}} \end{aligned} \tag{1}$$

Several properties of the Markov combination model can be easily established. By construction $(\psi_m, Y_m) \perp\!\!\!\perp (\psi_\ell, Y_\ell) \mid \phi$ for $m \neq \ell$ (see Figure 1). As noted above, the model (1) also has the attractive property that all (prior) marginal distributions are preserved: $p_{\text{comb}}(\phi, \psi_m, Y_m) = p_m(\phi, \psi_m, Y_m)$ for all $m = 1, \dots, M$. Massa and Lauritzen (2010) show that among the set of all distributions with this marginal preservation property, the distribution formed by Markov combination has maximal entropy and so can be viewed as the least constrained such distribution. The posterior distributions of ϕ and ψ_m , $m = 1, \dots, M$, under the Markov combination model account for all data, rather than just the submodel-specific data, and so naturally are not preserved.

3.3 Markov melding

Here we extend the framework above to the situation where the marginal distributions $p_1(\phi), \dots, p_M(\phi)$ of the link parameter differ by forming a *pooled density* $p_{\text{pool}}(\phi) = g(p_1(\phi), \dots, p_M(\phi))$ as a function g of the individual marginal densities. Here, and in what follows, we assume that such a pooled density exists; that g has been chosen such that $\int p_{\text{pool}}(\phi) d\phi = 1$; and that p_{pool} reflects an appropriate summary of the individual marginal distributions (we discuss options below).

Given $p_{\text{pool}}(\phi)$, we generalise Markov combination by replacing the common marginal distribution $p(\phi)$ for the link parameter in equation (1) with the pooled density:

$$\begin{aligned} p_{\text{meld}}(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M) &= p_{\text{pool}}(\phi) \prod_{m=1}^M p_m(\psi_m, Y_m \mid \phi) \\ &= p_{\text{pool}}(\phi) \prod_{m=1}^M \frac{p_m(\phi, \psi_m, Y_m)}{p_m(\phi)} \end{aligned} \tag{2}$$

We term this construction *Markov melding*. As with Markov combination, $(\psi_m, Y_m) \perp\!\!\!\perp (\psi_\ell, Y_\ell) \mid \phi$ for $m \neq \ell$ by construction. Similarly, $p_{\text{meld}}(\psi_m, Y_m \mid \phi) = p_m(\psi_m, Y_m \mid \phi)$ for all $m = 1, \dots, M$, that is, the submodel-specific conditional distributions, given the link parameter, are preserved in the joint model p_{meld} . The marginal distributions $p_m(\phi, \psi_m, Y_m)$ will, in general, not be preserved in the Markov melded model. However, we show in the next section that Markov melding can be viewed as Markov combination of altered versions of the submodels. The marginal distributions of the altered submodels will thus be preserved.

Once the new model $p_{\text{meld}}(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M)$ has been formed by Markov melding, posterior inference conditioning on the data $Y_1 = y_1, \dots, Y_M = y_M$ can be performed (see Section 4).

3.3.1 Markov combination as marginal replacement

Markov melding can be seen as the approach closest to Markov combination when the marginals $p_1(\phi), \dots, p_M(\phi)$ for the link parameter are not identical. To consider Markov combination in this context, we must first alter the original submodels $p_m(\phi, \psi_m, Y_m)$, $m = 1, \dots, M$ so that the marginals $p_1(\phi), \dots, p_M(\phi)$ for the link parameter are consistent (otherwise Markov combination is not defined). This can be achieved by *marginal replacement*, where a new model $p_{\text{repl},m}(\phi, \psi_m, Y_m)$ is formed by replacing the marginal distribution $p_m(\phi)$ of ϕ in the original model $p_m(\phi, \psi_m, Y_m)$ by a new marginal distribution

$p_{\text{new}}(\phi)$:

$$\begin{aligned} p_{\text{repl},m}(\phi, \psi_m, Y_m) &= p_m(\psi_m, Y_m \mid \phi) p_{\text{new}}(\phi) \\ &= \frac{p_m(\phi, \psi_m, Y_m)}{p_m(\phi)} p_{\text{new}}(\phi) \end{aligned} \quad (3)$$

By choosing $p_{\text{new}}(\phi) = p_{\text{pool}}(\phi)$, Markov melding (2) amounts to applying Markov combination (1) to submodels satisfying the consistency condition after marginal replacement via (3). Note that the (prior) marginals $p_{\text{meld}}(\phi, \psi_m, Y_m)$ in the Markov melded model match the marginals $p_{\text{repl},m}(\phi, \psi_m, Y_m)$ of the submodels after marginal replacement.

Each submodel $p_m(\phi, \psi_m, Y_m)$ could be altered to achieve a prescribed marginal $p_{\text{new}}(\phi)$ by methods other than the marginal replacement in (3) (we discuss one alternative in Supplementary Material D), but marginal replacement has several attractive properties. In particular, it can be interpreted as a generalisation of Bayesian updating in the light of new information (see Supplementary Material A). By extending a similar argument used in Poole and Raftery (2000), it can also be shown that $p_{\text{repl},m}$ minimises the Kullback-Leibler divergence D_{KL} of a distribution $q(\phi, \psi, Y)$ to $p_m(\phi, \psi_m, Y_m)$ under the constraint that the marginals on ϕ agree, $q(\phi) = p_{\text{new}}(\phi)$:

$$p_{\text{repl},m}(\phi, \psi_m, Y_m) = \operatorname{argmin}_q \{ D_{\text{KL}}(q \parallel p_m) \mid q(\phi) = p_{\text{new}}(\phi) \text{ for all } \phi \}$$

Details are provided in Supplementary Material A.

3.3.2 Marginal replacement for deterministic variables

Care is required when some of the dependencies in a submodel are deterministic. The considerations are identical to those for Bayesian melding (Poole and Raftery, 2000), where priors on the input and output of deterministic functions are combined. Specifically, assume the k -dimensional link parameter ϕ is deterministically related to a ℓ -dimensional parameter θ , $k \leq \ell$, in a model $p(\phi, \theta, \psi, Y)$, with Y denoting the observable random variables and ψ the remaining multivariate random variables. The probability model is effectively given by $p(\theta, \psi, Y)$ and ϕ follows an induced distribution. We assume ϕ is exclusively a deterministic function $\phi(\theta)$ of the random variable θ .

We need to ensure that it is possible to apply marginal replacement to $\phi(\theta)$. Specifically, we want the transformed variable $\phi(\theta)$ to assume a new marginal distribution $p_{\text{new}}(\phi)$. In order to apply marginal replacement, we must assume that $\phi(\theta)$ is an invertible function or, in the case of $k < \ell$, that $\phi(\theta)$ can be expanded into an invertible function $\phi_e(\theta) = (\phi(\theta), t(\theta))$. We denote the inverse function by $\theta(\phi, t)$. The function ϕ_e induces a probability distribution on (ϕ, t, ψ, Y) which can be represented as

$$p(\phi, t, \psi, Y) = p(\theta(\phi, t), \psi, Y) J_\theta(\phi, t)$$

where $J_\theta(\phi, t)$ is the Jacobian determinant for the transformation $\theta(\phi, t)$. The marginal distribution on ϕ can now be obtained as $p(\phi) = \int p(\phi, t, \psi, Y) dt d\psi dY$. We show in Supplementary Material B that this definition of the marginal distribution is independent of the chosen parametric extension $t(\theta)$ and that marginal replacement, as defined in equation (3), can be applied to the output of a deterministic function:

$$p_{\text{repl}}(\theta, \psi, Y) = \frac{p(\theta, \psi, Y)}{p(\phi(\theta))} p_{\text{new}}(\phi(\theta)) \quad (4)$$

Equation (4) justifies the approach we will take in the A/H1N1 example (Section 5.1), where ϕ is the output of a deterministic function.

3.3.3 Pooling marginal distributions

The pooling function g determines the marginal distributions $p_{\text{meld}}(\phi, \psi_m, Y_m)$, which, in general, will not match those in the original submodels. It must, therefore, be chosen subjectively, ensuring that the pooled density $p_{\text{pool}}(\phi)$ appropriately represents prior knowledge of the link parameter ϕ . Various standard pooling functions have been suggested in the multiple-expert elicitation literature (see, for example, Clemen and Winkler, 1999; O'Hagan *et al.*, 2006). The only difference here is that we propose to pool marginal prior distributions of submodels, rather than directly-specified priors. A simple option is *linear pooling*,

$$p_{\text{pool}}(\phi) = \frac{1}{K_{\text{lin}}(w)} \sum_{m=1}^M w_m p_m(\phi), \quad K_{\text{lin}}(w) = \int \sum_{m=1}^M w_m p_m(\phi) d\phi$$

where $w = (w_m)_{m=1}^M$, with $w_m \geq 0$ to weight the submodel priors. An alternative is *log pooling*,

$$p_{\text{pool}}(\phi) = \frac{1}{K_{\text{log}}(w)} \prod_{m=1}^M p_m(\phi)^{w_m}, \quad K_{\text{log}}(w) = \int \prod_{m=1}^M p_m(\phi)^{w_m} d\phi$$

with $w_m \geq 0$, a logarithmic version of the linear pooling. A special case of log pooling is *product of experts (PoE) pooling* (Hinton, 2002) when $w_m = 1$ for all m

$$p_{\text{pool}}(\phi) = \frac{1}{K_{\text{poe}}} \prod_{m=1}^M p_m(\phi), \quad K_{\text{poe}} = \int \prod_{m=1}^M p_m(\phi) d\phi$$

in which equal weight is given to each submodel prior. A further special case of linear or log pooling is *dictatorial pooling* $p_{\text{pool}}(\phi) = p_{m_0}(\phi)$ when one submodel m_0 is considered authoritative. Dictatorial pooling corresponds to left (or right) composition in the terminology of Massa and Lauritzen (2010). When using log pooling or PoE it is necessary that there is a region on which all submodels have support.

Figure 4 shows the pooled density when combining two normal distributions under three different pooling rules with three choices of weights. The PoE approach is arguably the least intuitive pooling function due to the rather concentrated combined distribution implied. However, the required computation is greatly simplified (see Section 4), making PoE pooling an attractive option for this reason.

3.4 Splitting models with Markov melding

Suppose we wish to split a large joint model $p(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M)$ into M submodels $p_m(\phi, \psi_m, Y_m)$, $m = 1, \dots, M$ in such a way that joining the submodels using Markov melding recovers the original, joint model. If $(\psi_m, Y_m) \perp\!\!\!\perp (\psi_\ell, Y_\ell) \mid \phi$ for $m \neq \ell$ in the original joint model, then suitable submodels are

$$p_m(\phi, \psi_m, Y_m) = p(\psi_m, Y_m \mid \phi) p_m(\phi), \quad m = 1, \dots, M$$

where $p_1(\phi), \dots, p_M(\phi)$ are dummy marginal distributions, specified so that the pooled distribution $p_{\text{pool}}(\phi) = g(p_1(\phi), \dots, p_M(\phi))$ is the same as the original marginal distribution $p(\phi)$ for a chosen pooling function g . In this context, the choice of g and marginal distribution $p_m(\phi)$ in each submodel m does not represent a subjective modelling judgement, provided the pooled distribution $p_{\text{pool}}(\phi)$ matches the original prior $p(\phi)$. The

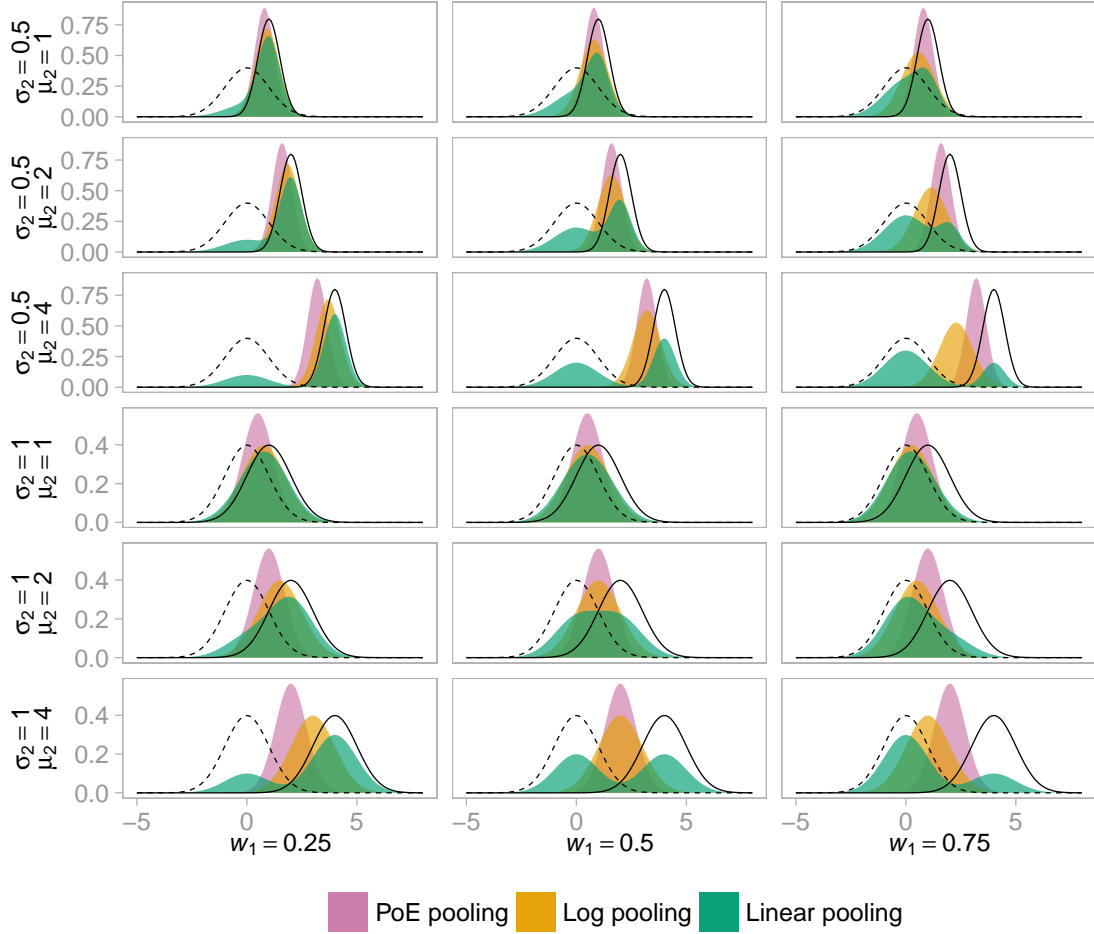


Figure 4: Pooled densities under PoE, log and linear pooling, with $w_1 = 0.25, 0.5$ and 0.75 (and $w_2 = 1 - w_1$), formed by pooling a $N(0, 1)$ density (---) and a $N(\mu_2, \sigma_2^2)$ density (—) with $\mu_2 = 1, 2, 4$, and $\sigma_2 = 0.5, 1$.

choice can be made freely to enable efficient computation. A vanilla choice is to set $p_m(\phi) = p(\phi)^{1/M}$, and use PoE pooling, but other choices are available. For example, PoE pooling is suitable for *any* factorisation of $p(\phi)$ into M factors.

It is important to stress that a splitting strategy based on Markov melding is suitable *only* if $(\psi_m, Y_m) \perp\!\!\!\perp (\psi_\ell, Y_\ell) \mid \phi$ for $m \neq \ell$, that is, conditioning on the link variable ϕ renders the parts intended for splitting conditionally independent.

Figure 5 shows a few stylised situations with $M = 2$ where splitting for computational purposes might be desirable. The joint distributions for all models is $p(\phi, \psi_1, \psi_2, Y_1, Y_2)$. The model in Figure 5a can be split into $p_1(\phi, \psi_1, Y_1) = p(\phi, Y_1 \mid \psi_1)p(\psi_1)$ and $p_2(\phi, \psi_2, Y_2) = p(\psi_2, Y_2 \mid \phi)p_2(\phi)$, with dummy prior distribution $p_2(\phi)$, which could be different and computationally simpler than $p(\phi) = \int p(\phi, Y_1 \mid \psi_1)p(\psi_1) d\psi_1 dY_1$. Markov melding, with dictatorial pooling $p_{\text{pool}}(\phi) = p_1(\phi)$, results in

$$\begin{aligned} p_{\text{meld}}(\phi, \psi_1, \psi_2, Y_1, Y_2) &= p_1(\phi) \frac{p_1(\phi, \psi_1, Y_1)}{p_1(\phi)} \frac{p_2(\phi, \psi_2, Y_2)}{p_2(\phi)} \\ &= p(\phi, Y_1 \mid \psi_1) p(\psi_1) p_2(\psi_2, Y_2 \mid \phi) = p(\phi, \psi_1, \psi_2, Y_1, Y_2), \end{aligned}$$

leading the original, joint model.

The case in Figure 5b is similar to the example in Section 2.2 (see Section 3.5.2 for a definition of splitting in this case). Note that in Figures 5a and 5b, the dependencies

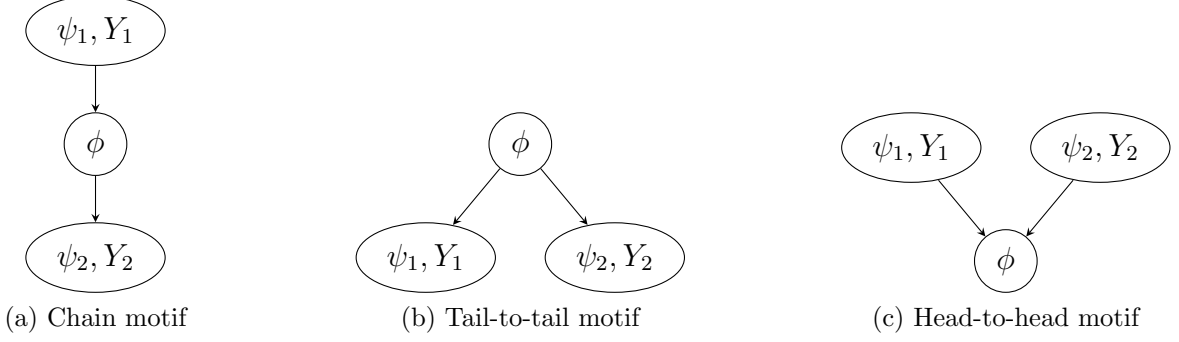


Figure 5: DAG representations of stylised situations where model splitting might be desirable. Splitting the joint model is possible in (a) and (b), but not in (c).

between the nodes can include deterministic (logical) dependence, provided $(\psi_1, Y_1) \perp\!\!\!\perp (\psi_2, Y_2) \mid \phi$, as usual.

The case in Figure 5c cannot be split into $p_1(\phi, \psi_1, Y_1)$ and $p_2(\phi, \psi_2, Y_2)$ by Markov melding model splitting because $(\psi_1, Y_1) \not\perp\!\!\!\perp (\psi_2, Y_2) \mid \phi$.

3.5 Markov melding in the motivating examples

3.5.1 Joining: A/H1N1 influenza evidence synthesis

Applying Markov melding to the evidence synthesis example (Section 2.1) involves joining the ICU submodel (Figure 2a) with density $p_1(\phi, \theta, \pi^{\text{pos}}, Y)$, and the severity submodel (Figure 2b) with density $p_2(\phi, \chi, \pi^{\text{det}})$. With pooled density $p_{\text{pool}}(\phi)$ for the link variable ϕ , Markov melding results in

$$p_{\text{meld}}(\phi, \theta, \pi^{\text{pos}}, \chi, \pi^{\text{det}}, Y) = p_{\text{pool}}(\phi) p_1(\theta, \pi^{\text{pos}}, Y \mid \phi) p_2(\chi, \pi^{\text{det}} \mid \phi)$$

where $Y_1 = Y$, $\psi_1 = \{\theta, \pi^{\text{pos}}\}$, $Y_2 = \emptyset$ and $\psi_2 = \{\chi, \pi^{\text{det}}\}$ in the notation of equation (2).

3.5.2 Splitting: large ecology model

The original, joint model (Figure 3a) in Section 2.2, with density $p(\phi, \lambda, \gamma, Y, X)$, can be split into separate submodels (Figure 3b) with densities $p_1(\phi, \lambda, Y)$ and $p_2(\phi, \gamma, X)$, provided the priors $p_1(\phi)$ and $p_2(\phi)$ in the separate submodels are such that $p_{\text{pool}}(\phi) = g(p_1(\phi), p_2(\phi))$ equals the original marginal distribution $p(\phi)$, for some choice of pooling function g . Clearly, Markov melding the submodels recovers the joint model,

$$\begin{aligned} p_{\text{meld}}(\phi, \lambda, \gamma, Y, X) &= p_{\text{pool}}(\phi) p_1(\lambda, Y \mid \phi) p_2(\gamma, X \mid \phi) \\ &= p(Y \mid \lambda, \phi) p(\lambda) p(X \mid \gamma, \phi) p(\gamma) p(\phi), \end{aligned}$$

with $\psi_1 = \lambda$, $Y_1 = Y$, $\psi_2 = \gamma$, $Y_2 = X$ in the notation of equation (2).

4 Inference and computation

The joint posterior distribution, given data $Y_m = y_m$, $m = 1, \dots, M$, under the Markov melded model in equation (2) is

$$p_{\text{meld}}(\phi, \psi_1, \dots, \psi_M \mid y_1, \dots, y_M) \propto p_{\text{pool}}(\phi) \prod_{m=1}^M \frac{p_m(\phi, \psi_m, y_m)}{p_m(\phi)} \quad (5)$$

The degree of difficulty of inference for this posterior distribution depends on the specification of the constituent submodels. Our focus is settings in which, considered separately, each of the original collection of submodels is amenable to inference by standard Monte Carlo methods (for example, Metropolis *et al.*, 1953; Hastings, 1970; Neal, 2003).

In Section 4.1 we first consider a standard Metropolis-within-Gibbs sampler, but when the constituent submodels are complex, this sampler may be cumbersome and slow. We thus propose a multi-stage Metropolis-within-Gibbs sampler, in which inference for the full Markov melded model is generated iteratively in stages, starting with standard inference on one of the constituent submodels. This latter approach enables a convenient modular approach to inference. Both approaches, in general, require the marginal prior densities $p_m(\phi)$ of the link parameter under each model, $m = 1, \dots, M$, which will not usually be analytically tractable. In Section 4.2 we discuss approaches to estimating these densities, although there is no need to estimate them if PoE pooling is chosen. In Section 4.3 we show how approximate approaches, such as those used by Presanis *et al.* (2014), relate to the Markov melded model.

4.1 Metropolis-Hastings samplers

A Metropolis-Hastings sampler for the posterior distribution (5) can be constructed in the usual way. Candidate values $(\phi^*, \psi_1^*, \dots, \psi_M^*)$ for each parameter of the Markov melded model are drawn from a proposal distribution $q(\phi^*, \psi_1^*, \dots, \psi_M^* \mid \phi, \psi_1, \dots, \psi_M)$, based on the current values $(\phi, \psi_1, \dots, \psi_M)$ of the Markov chain. The candidate values are accepted with probability $\min(1, r)$, where r is in the form

$$r = \frac{R(\phi^*, \psi_1^*, \dots, \psi_M^*, \phi, \psi_1, \dots, \psi_M)}{R(\phi, \psi_1, \dots, \psi_M, \phi^*, \psi_1^*, \dots, \psi_M^*)}$$

where the *target-to-proposal density ratio* is

$$\begin{aligned} & R(\phi^*, \psi_1^*, \dots, \psi_M^*, \phi, \psi_1, \dots, \psi_M) \\ &= p_{\text{pool}}(\phi^*) \prod_{m=1}^M \frac{p_m(\phi^*, \psi_m^*, y_m)}{p_m(\phi^*)} \times \frac{1}{q(\phi^*, \psi_1^*, \dots, \psi_M^* \mid \phi, \psi_1, \dots, \psi_M)} \end{aligned}$$

4.1.1 Metropolis-within-Gibbs sampler

In a Metropolis-within-Gibbs sampling scheme (Müller, 1991), samples are drawn from the full conditional distribution of each latent parameter ψ_1, \dots, ψ_M , and then the link parameter ϕ in turn.

Latent parameter updates Markov melding does not introduce any extra complexities in sampling the parameters ψ_m in each submodel $m = 1, \dots, M$ (conditional on the link parameter ϕ) beyond those inherent to the original submodels, considered separately. Typically, they can be sampled using standard algorithms. For instance, a Metropolis-Hastings algorithm, in which we draw a candidate value ψ_m^* from a proposal distribution $q(\psi_m^* \mid \psi_m)$ based upon the current value ψ_m , will be feasible whenever the corresponding algorithm is feasible for estimation of the posterior distribution of the m^{th} submodel alone. In this case, the target-to-proposal density ratio simplifies to

$$R(\phi, \psi_1, \dots, \psi_m^*, \dots, \psi_M, \phi, \psi_1, \dots, \psi_M) = p_m(\phi, \psi_m^*, y_m) \times \frac{1}{q(\psi_m^* \mid \psi_m)}$$

since terms involving marginal densities for the link parameter ϕ in equation (5) cancel. This target-to-proposal density ratio is identical to that required for a Metropolis-Hastings update for the parameter ψ_m , conditional on the link parameter ϕ , when the m^{th} submodel alone is the target distribution.

Link parameter updates To update the link parameters, a candidate value ϕ^* is drawn from an appropriate proposal distribution $q(\phi^* | \phi)$, based upon the current value ϕ , and is accepted according to the target-to-proposal density ratio

$$R(\phi^*, \psi_1, \dots, \psi_M, \phi, \psi_1, \dots, \psi_M) = p_{\text{pool}}(\phi^*) \prod_{m=1}^M \frac{p_m(\phi^*, \psi_m, y_m)}{p_m(\phi^*)} \times \frac{1}{q(\phi^* | \phi)}$$

and rejected otherwise.

When the marginal distributions $p_m(\phi)$ or $p_{\text{pool}}(\phi)$ are not analytically tractable, we propose to use an approximation $\hat{p}_m(\phi)$ in their place, calculated using the methods described in Section 4.2. Note that, under PoE pooling, the terms involving the marginal distributions for the link parameter ϕ cancel, leaving the simplified target-to-proposal density ratio

$$R(\phi^*, \psi_1, \dots, \psi_M, \phi, \psi_1, \dots, \psi_M) = \prod_{m=1}^M p_m(\phi^*, \psi_m, y_m) \times \frac{1}{q(\phi^* | \phi)}$$

removing the need to estimate the marginal prior distribution for the link parameter ϕ .

4.1.2 Multi-stage Metropolis-within-Gibbs sampler

A multi-stage, alternative approach is computationally preferable when the constituent submodels are complex. This generalises the two stage approach in Lunn *et al.* (2013a). We assume a factorisation of the pooled prior $p_{\text{pool}}(\phi) = \prod_{m=1}^M p_{\text{pool},m}(\phi)$ for ϕ , in any suitable manner that enables efficient sampling. A default factorisation for any pooling function sets $p_{\text{pool},m}(\phi) = p_{\text{pool}}(\phi)^{1/M}$, but when PoE pooling is used the factorisation with $p_{\text{pool},m}(\phi) = p_m(\phi)$ is more computationally efficient, as we describe below. The aim then is to sample from

$$p_{\text{meld},\ell}(\phi, \psi_1, \dots, \psi_\ell | y_1, \dots, y_\ell) \propto \prod_{m=1}^{\ell} \left(\frac{p_m(\phi, \psi_m, y_m)}{p_m(\phi)} p_{\text{pool},m}(\phi) \right) \quad (6)$$

iteratively in stages $\ell = 1, \dots, M$. Note that $p_{\text{meld},M}(\phi, \psi_1, \dots, \psi_M | y_1, \dots, y_M) = p_{\text{meld}}(\phi, \psi_1, \dots, \psi_M | y_1, \dots, y_M)$ and so after M stages, the samples obtained reflect the posterior distribution (5) of the full Markov melded model. Note that each $p_m(\phi)$ (and thus also $p_{\text{pool}}(\phi)$) can be estimated from the submodels in advance and independently of the following sampling scheme, as we describe in Section 4.2.

Stage 1. We obtain H_1 samples $(\phi^{(h,1)}, \psi_1^{(h,1)})$, $h = 1, \dots, H_1$, drawn from $p_{\text{meld},1}(\phi, \psi_1 | y_1)$. The most appropriate method for obtaining such samples depends on the nature of the submodel $p_1(\phi, \psi_1, Y_1)$; typically, standard Monte Carlo methods, such as MCMC, will be suitable.

Stage ℓ . After we have sampled up to stage $\ell - 1$ from (6), we construct a Metropolis-within-Gibbs sampler for stage ℓ for the parameters $(\phi, \psi_1, \dots, \psi_\ell)$ given data (y_1, \dots, y_ℓ) . The parameter ψ_ℓ is updated, conditional on the link parameter ϕ and parameters $\psi_1, \dots, \psi_{\ell-1}$ using a standard algorithm, such as a Metropolis-Hastings sampler with a target-to-proposal density ratio

$$R(\phi, \psi_1, \dots, \psi_\ell^*, \phi, \psi_1, \dots, \psi_\ell) = p_\ell(\phi, \psi_\ell^*, y_\ell) \times \frac{1}{q(\psi_\ell^* | \phi, \psi_\ell)}$$

For updating the parameters ψ_1, \dots, ψ_ℓ and the link parameter ϕ we use the samples from stage $\ell - 1$ as a proposal distribution. Specifically, we draw an index d uniformly at random from $\{1, \dots, H_{\ell-1}\}$, so that

$$(\phi^*, \psi_1^*, \dots, \psi_{\ell-1}^*) = (\phi^{(d, \ell-1)}, \psi_1^{(d, \ell-1)}, \dots, \psi_{\ell-1}^{(d, \ell-1)}) \sim p_{\text{meld}, \ell-1}(\phi, \psi_1, \dots, \psi_{\ell-1} | y_1, \dots, y_{\ell-1})$$

The attraction of this particular proposal distribution is the resulting cancellation of likelihood terms for the first $\ell - 1$ submodels in the target-to-proposal density ratio, meaning this update step can be performed quickly

$$R(\phi^*, \psi_1^*, \dots, \psi_{\ell-1}^*, \psi_\ell, \phi, \psi_1, \dots, \psi_{\ell-1}, \psi_\ell) = \frac{p_\ell(\phi^*, \psi_\ell, y_\ell)}{p_\ell(\phi^*)} p_{\text{pool}, \ell}(\phi^*) \quad (7)$$

Once sampling for this stage has converged, H_ℓ samples $(\phi^{(h, \ell)}, \psi_1^{(h, \ell)}, \dots, \psi_\ell^{(h, \ell)})$, $h = 1, \dots, H_\ell$, are obtained for the next stage.

The density ratio (7) does not depend on parameters $\psi_1, \dots, \psi_{\ell-1}$, so if interest focuses entirely on the parameters (ϕ, ψ_ℓ) then $\psi_1, \dots, \psi_{\ell-1}$ can be ignored in stage ℓ of the multi-stage sampling: they do not need to be monitored or updated by the sampling algorithm. The multi-stage sampler is nevertheless still sampling from the joint target distribution $p_{\text{meld}, \ell}(\phi, \psi_1, \dots, \psi_\ell | y_1, \dots, y_\ell)$. Stage ℓ is influencing the acceptance or rejection of samples of $p_{\text{meld}, \ell-1}(\phi, \psi_1, \dots, \psi_{\ell-1} | y_1, \dots, y_{\ell-1})$ from the previous stage, thus adjusting this distribution according to the requirements of the joint model.

In general, evaluation of ratio (7) requires estimates of the marginal distribution of the link parameter under the ℓ^{th} submodel, which can be obtained as described in Section 4.2. However, if PoE pooling is used to pool marginals on ϕ the ratio simplifies to $R(\phi^*, \psi_1^*, \dots, \psi_{\ell-1}^*, \psi_\ell, \phi, \psi_1, \dots, \psi_{\ell-1}, \psi_\ell) = p_\ell(\phi^*, \psi_\ell, y_\ell)$, meaning that no estimates of the marginal distribution are required.

4.2 Estimating marginal distributions

The marginal densities $p_m(\phi)$ of the link parameter under each of the M submodels are central to Markov melding, and in particular are required to evaluate the acceptance probability of proposals within the MCMC samplers we proposed above. However, these marginals are not generally analytically tractable, except when the prior distribution $p_m(\phi)$ is directly-specified as a standard, tractable distribution, such as when ϕ appears as a founder node in a DAG representation of the submodel. When not available analytically, we can estimate the marginal density $p_m(\phi)$ for each submodel m by kernel density estimation (Henderson and Parmeter, 2015) with samples drawn from $p_m(\phi) = \iint p_m(\phi, \psi_m, Y_m) d\psi_m dY_m$ by standard (forward) Monte Carlo.

4.3 Normal two-stage approximation method

Approximate approaches for joining submodels are widely used in applied research. In this section, we show that approximate inference for the Markov melded model formed by joining submodels (with PoE pooling) can be produced using a standard normal approximation approach.

Consider the case when the Markov melded model $p_{\text{meld}}(\phi, \psi_1, \psi_2, Y_1, Y_2)$ is formed by joining $M = 2$ submodels $p_1(\phi, \psi_1, Y_1)$ and $p_2(\phi, \psi_2, Y_2)$. Suppose that ψ_1 is not a parameter of interest in the posterior distribution, so that it can be integrated over

$$\begin{aligned} p_{\text{meld}}(\phi, \psi_2, Y_1, Y_2) &= \int p_{\text{meld}}(\phi, \psi_1, \psi_2, Y_1, Y_2) d\psi_1 \\ &= p_{\text{pool}}(\phi) \int p_1(\psi_1, Y_1 | \phi) d\psi_1 p_2(\psi_2, Y_2 | \phi) \\ &= p_{\text{pool}}(\phi) p(Y_1 | \phi) p_2(\psi_2, Y_2 | \phi) \end{aligned} \quad (8)$$

An approximate two stage sampler that mimics the multi-stage sampler (above) can then be constructed for this marginal distribution.

Stage 1 Fit the model $p_1(\phi, \psi_1, Y_1)$ to obtain posterior samples from $p_1(\phi | Y_1)$, and approximate by a (multivariate) normal distribution with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$.

$$p_1(\phi | Y_1) \approx p_N(\phi | \hat{\mu}, \hat{\Sigma}) = p_N(\hat{\mu} | \phi, \hat{\Sigma})$$

Stage 2 Since $p_1(\phi, Y_1) \propto p_1(\phi | Y_1) \approx p_N(\hat{\mu} | \phi, \hat{\Sigma})$, we obtain an approximation for (8) by replacing $p_1(\phi | Y_1)$ by $p_N(\hat{\mu} | \phi, \hat{\Sigma})$.

$$\begin{aligned} p_{\text{meld}}(\phi, \psi_2, Y_1, Y_2) &\propto p_{\text{pool}}(\phi) \frac{p_1(\phi | Y_1) p_2(\phi, \psi_2, Y_2)}{p_1(\phi)} \\ &\approx p_{\text{pool}}(\phi) \frac{p_N(\hat{\mu} | \phi, \hat{\Sigma}) p_2(\phi, \psi_2, Y_2)}{p_1(\phi)} \end{aligned}$$

This two-stage approximate approach is commonly used in practice (see Section 6) in the form $p_{\text{meld}}(\phi, \psi_2, Y_1, Y_2) \approx c p_N(\hat{\mu} | \phi, \hat{\Sigma}) p_2(\phi, \psi_2, Y_2)$, with c a data dependent constant. In this case the likelihood of the second model is modified by a factor $p_N(\hat{\mu} | \phi, \hat{\Sigma})$ (in a DAG representation a dependency of the constant $\hat{\mu}$ on ϕ and the constant $\hat{\Sigma}$ is added). This approach can be viewed as approximate Markov melding with PoE pooling, in which one model is represented by a normal approximation.

If, instead of PoE pooling, one wishes to regard the marginal $p_2(\phi)$ on the link variable ϕ as authoritative and thus fully retain it, dictatorial pooling $p_{\text{pool}}(\phi) = p_2(\phi)$ leads to the variant

$$\begin{aligned} p_{\text{meld}}(\phi, \psi_2, Y_1, Y_2) &\propto \frac{p_1(\phi | Y_1)}{p_1(\phi)} p_2(\phi, \psi_2, Y_2) \approx \frac{p_N(\phi | \hat{\mu}, \hat{\Sigma})}{p_N(\phi | \hat{\mu}_0, \hat{\Sigma}_0)} p_2(\phi, \psi_2, Y_2) \\ &\propto p_N(\phi | \mu_c, \Sigma_c) p_2(\phi, \psi_2, Y_2) \end{aligned}$$

where $\hat{\mu}_0$ and $\hat{\Sigma}_0$ are an estimate of the mean and covariance of the prior marginal $p_1(\phi)$, which can be obtained at stage one in parallel to the posterior by sampling from the prior model, and

$$\Sigma_c^2 = \left(\hat{\Sigma}^{-1} - \hat{\Sigma}_0^{-1} \right)^{-1}, \quad \mu_c = \Sigma_c \left(\hat{\Sigma}^{-1} \hat{\mu} - \hat{\Sigma}_0^{-1} \hat{\mu}_0 \right).$$

Adjusting according to $\hat{\mu}_0$ and $\hat{\Sigma}_0$ removes the prior $p_1(\phi)$ from approximate joint model.

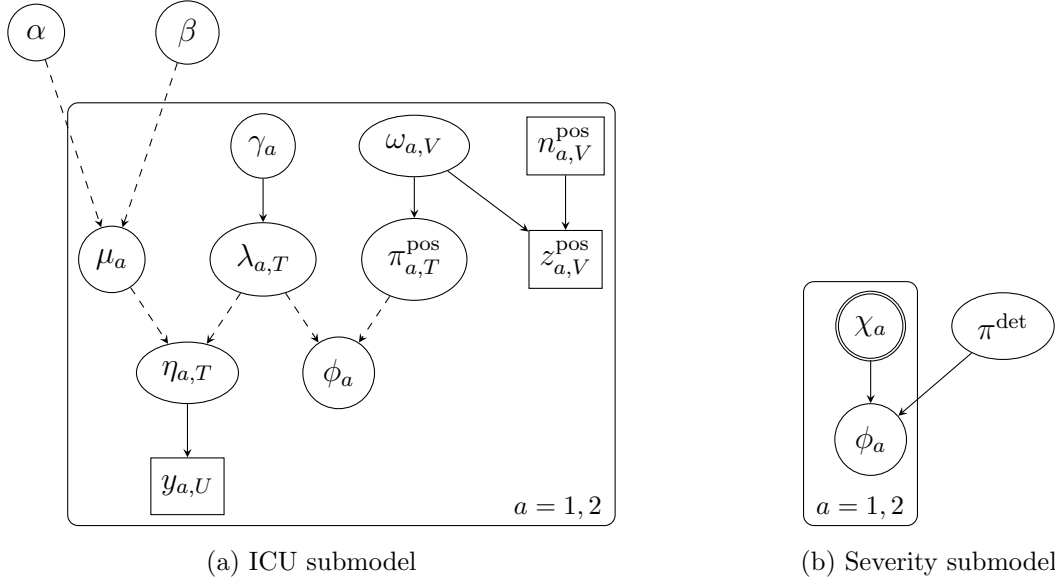


Figure 6: DAG representations of the submodels of A/H1N1 influenza. Repeated variables are enclosed by a rounded rectangle, with the label denoting the range of repetition (Buntine, 1994). For simplicity the time domain is suppressed: parameters with subscripts T , U and V are collections of parameters across the time range denoted by the subscript. For example, $y_{a,U} = \{y_{a,t} : t \in U\}$.

5 Results

5.1 Joining: A/H1N1 influenza evidence synthesis

Figure 6 shows DAG representations of the two submodels outlined in Section 2.1.

5.1.1 ICU submodel

The main data source in the ICU submodel is prevalence-type data from the Department of Health’s Winter Watch scheme (Department of Health, 2011), which records the total number of patients with *suspected* pandemic A/H1N1 influenza infection across all ICUs in England. Weekly observations are available between the 15th December 2010 (day $t = 8$) and 23rd February 2011 ($t = 78$). Denote by $y_{a,t}$ the observation on day $t \in U = \{8, 15, 22, \dots, 78\}$ for age group $a \in \{1, 2\}$, where $a = 1$ indicates a child under 17 years old and $a = 2$ indicates an adult. We model the system of ICU admission and discharges, deaths or transfers out of ICU as an immigration-death process, assuming new ICU admissions follow an inhomogeneous Poisson process with rate $\lambda_{a,t}$ at time t , and the length of stay in ICU is exponentially distributed with rate μ_a . Then the number of patients admitted up to time t who are still present in ICU at time t follows a thinned inhomogeneous Poisson process and the observed number of prevalent patients $y_{a,t} \sim \text{Po}(\eta_{a,t})$, $a \in \{1, 2\}$, $t \in U$, with expectation, under a discretised formulation with daily time steps, given by $\eta_{a,t} = \sum_{u=1}^t \lambda_{a,u} \exp\{-\mu_a(t-u)\}$, $t \in T = \{1, \dots, 78\}$. We assume $\eta_{a,1} = 0$ to enforce the assumption that no patients with suspected ‘flu were in ICU a week before observations began.

The product of the admissions rate $\lambda_{a,t}$ and the positivity rate $\pi_{a,t}^{\text{pos}}$ gives the expected number of *confirmed* (as opposed to *suspected*) new admissions in age group a on day t .

Our link parameter $\phi = (\phi_1, \phi_2)$ is the sum of these quantities over time:

$$\phi_a = \sum_{t \in T} \pi_{a,t}^{\text{pos}} \lambda_{a,t}, \quad a = 1, 2.$$

We model the positivity rate $\pi_{a,t}^{\text{pos}}$ using weekly virological positivity data from the sentinel laboratory surveillance system Data Mart (Public Health England, 2014), which records the number $z_{a,v}^{\text{pos}}$ of A/H1N1-positive swabs out of the total number $n_{a,v}^{\text{pos}}$ tested during week $v \in V = \{1, \dots, 11\}$ in age group $a \in \{1, 2\}$ (note that the adult age group includes 16 year olds, in contrast to Winter Watch). We assume a uniform prior $\pi_{a,t}^{\text{pos}} \sim \text{Unif}(\omega_{a,v}, 1)$, $t \in T$, for the true positivity, where $v = 1$ for $t = 1, \dots, 14$ and $v = \lfloor (t - 1)/7 \rfloor$ for $t = 15, \dots, 78$, and where the lower bound $\omega_{a,v}$ is informed by a binomial model for the positivity data: $z_{a,v}^{\text{pos}} \sim \text{Bin}(n_{a,v}^{\text{pos}}, \omega_{a,v})$, $v \in V$. For the admission rates $\lambda_{a,t}$, Presanis *et al.* (2014) assumed a prior with undefined variance, but this renders the methods in Section 4.2 inappropriate for estimation of the marginal distribution $p(\phi)$. Instead, we assume a random-walk prior for the admission rates with $\log(\lambda_{a,1}) \sim \text{Unif}(0, 250)$ and $\log(\lambda_{a,t}) \sim \text{N}(\log(\lambda_{a,t-1}), \gamma_a^{-2})$ for $t = 2, \dots, 78$, with $\gamma_a \sim \text{Unif}(0.1, 2.7)$. For the length of ICU stays we assume constant age-group specific rates $\mu_1 = \exp(-\alpha)$ and $\mu_2 = \exp(-\{\alpha + \beta\})$, with $\alpha \sim \text{N}(2.7058, 0.0788^2)$ and $\beta \sim \text{N}(-0.4969, 0.2048^2)$, where the mean and variances are chosen to match those indicated by individual-level hospitalisation data (Presanis *et al.*, 2014).

5.1.2 Severity submodel

We consider a simplified version of the full, complex severity submodel in Presanis *et al.* (2014). The Winter Watch ICU data are only available for a portion of the time of the ‘third wave’ of the A/H1N1 pandemic, and so the cumulative number of confirmed new admissions ϕ_a from the ICU is a lower bound for the true number χ_a of ICU admissions during the third wave. We thus assume $\phi_a \sim \text{Bin}(\chi_a, \pi^{\text{det}})$, $a \in \{1, 2\}$, where π^{det} is the age-constant detection probability, to which we assign a Beta(6, 4) prior. We incorporate the remaining evidence in the full severity submodel of Presanis *et al.* (2014) via informative priors $\chi_1 \sim \text{Lognormal}(4.93, 0.17^2)$ and $\chi_2 \sim \text{Lognormal}(7.71, 0.23^2)$.

5.1.3 Markov melded model

We joined the submodels by Markov melding as described in Section 3.5.1. We considered linear and log pooling with pooling weight $w_1 = 0.25, 0.5$ and 0.75 (and $w_2 = 1 - w_1$), as well as PoE pooling.

We estimated the marginal priors for $\phi = (\phi_1, \phi_2)$ under the ICU and severity submodels using kernel density estimation with a bivariate t -distribution kernel, using 5×10^4 independent draws sampled from the corresponding submodel by forward Monte Carlo. The marginal priors are shown in Figure 7a. Note that the ICU submodel prior for ϕ is extremely flat, whereas the severity submodel prior is concentrated on a small part of the parameter space. The combined density using each of the pooling rules (with $w_1 = w_2 = 0.5$) is shown in Figure 7b. Linear and PoE pooling in this case lead to similar densities, whereas the log pool prior is more dispersed.

We then estimated, in stage one, the posterior distribution of the link parameter ϕ under the ICU submodel alone. We drew 5 million iterations from the ICU submodel using JAGS (Plummer, 2015b), retaining every 100th iteration after discarding 5×10^4 iterations as burn-in. In stage two, for the Markov melded models under linear, log and

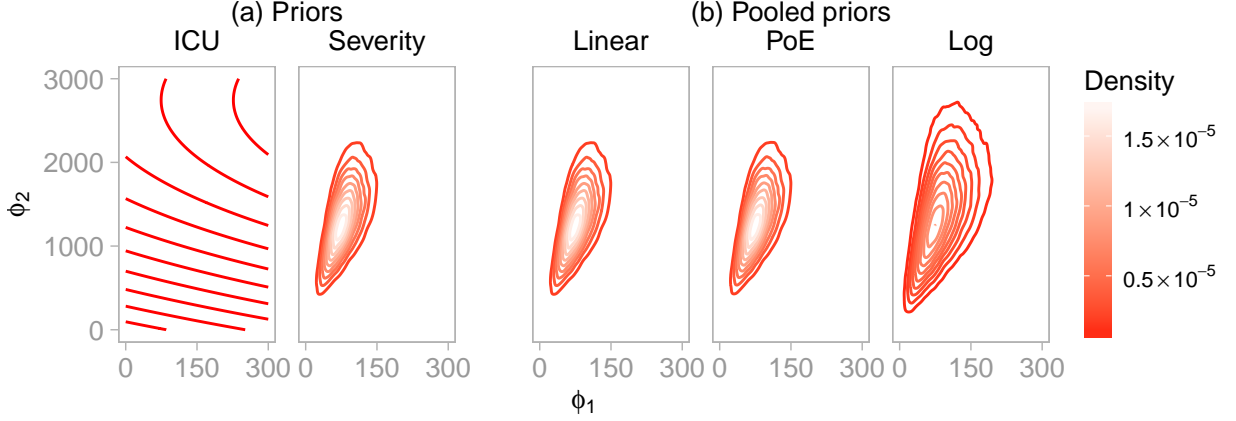


Figure 7: Prior distributions for ϕ_a , the cumulative number of *confirmed* new admissions in age group a , in the A/H1N1 influenza evidence synthesis: (a) under the ICU and severity submodels; (b) pooled priors under three pooling functions with $w_1 = w_2 = 0.5$.

PoE pooling, we drew 5×10^5 samples using the multi-stage Metropolis-within-Gibbs sampler, with the first 10^4 samples discarded as burn-in.

Figure 8 shows the posterior distribution under these three prior assumptions, and using the normal approximation approach (fitted using OpenBUGS). There is considerable agreement between the various approaches, except the normal approximation, which does not reflect the long upper tail of the parameters (ϕ_1 and χ_1) relating to the child age group. This is because the posterior distribution of ϕ_1 under the ICU submodel is positively skewed, meaning that the normal approximation is a relatively poor fit in this age group. The Markov melding results appear to be robust to the choice of pooling rule and pooling weight w_1 in this example: the likelihood from the ICU submodel appears to dominate over the pooled prior in each choice.

5.2 Splitting: large ecology model

Figure 9 is a DAG representation of the full, joint model outlined in Section 2.2.

5.2.1 Mark-recapture-recovery data

Mark-recapture-recovery data y_{t_1, t_2} record the number of ringed birds released before May in year $t_1 = 1, \dots, 35$, and recovered (dead) in the 12 months up to April in year $t_2 = t_1 + 1, \dots, 36$. The years correspond to observations for releases from 1963 ($t = 1$) to 1997 and recoveries from 1964 to 1998. The number of birds $y_{t_1, 37}$ released in year t_1 and never recovered is also available. We assume

$$(y_{t_1, t_1+1}, \dots, y_{t_1, 37}) \sim \text{Mult}(\pi_{t_1, t_1+1}, \dots, \pi_{t_1, 37}), \quad t_1 = 1, \dots, 35.$$

We model the probability π_{t_1, t_2} of recovery in year t_2 following release in year t_1 in terms of the recovery rate λ_t , and the survival rates $\eta_{C, t}$ and $\eta_{A, t}$ for immature (1 year old) and breeding (2 years or older) birds, respectively, up to April of year t :

$$\pi_{t_1, t_2} = \begin{cases} \lambda_{t_2}(1 - \eta_{C, t_2}) & t_1 = 1, \dots, 35, t_2 = t_1 + 1 \\ \lambda_{t_2}\eta_{C, t_1+1}(1 - \eta_{A, t_2}) & t_1 = 1, \dots, 34, t_2 = t_1 + 2 \\ \lambda_{t_2}\eta_{C, t_1+1}(1 - \eta_{A, t_2}) \prod_{u=t_1+2}^{t_2-1} \eta_{A, u} & t_1 = 1, \dots, 33, t_2 = t_1 + 3, \dots, 36 \end{cases}$$

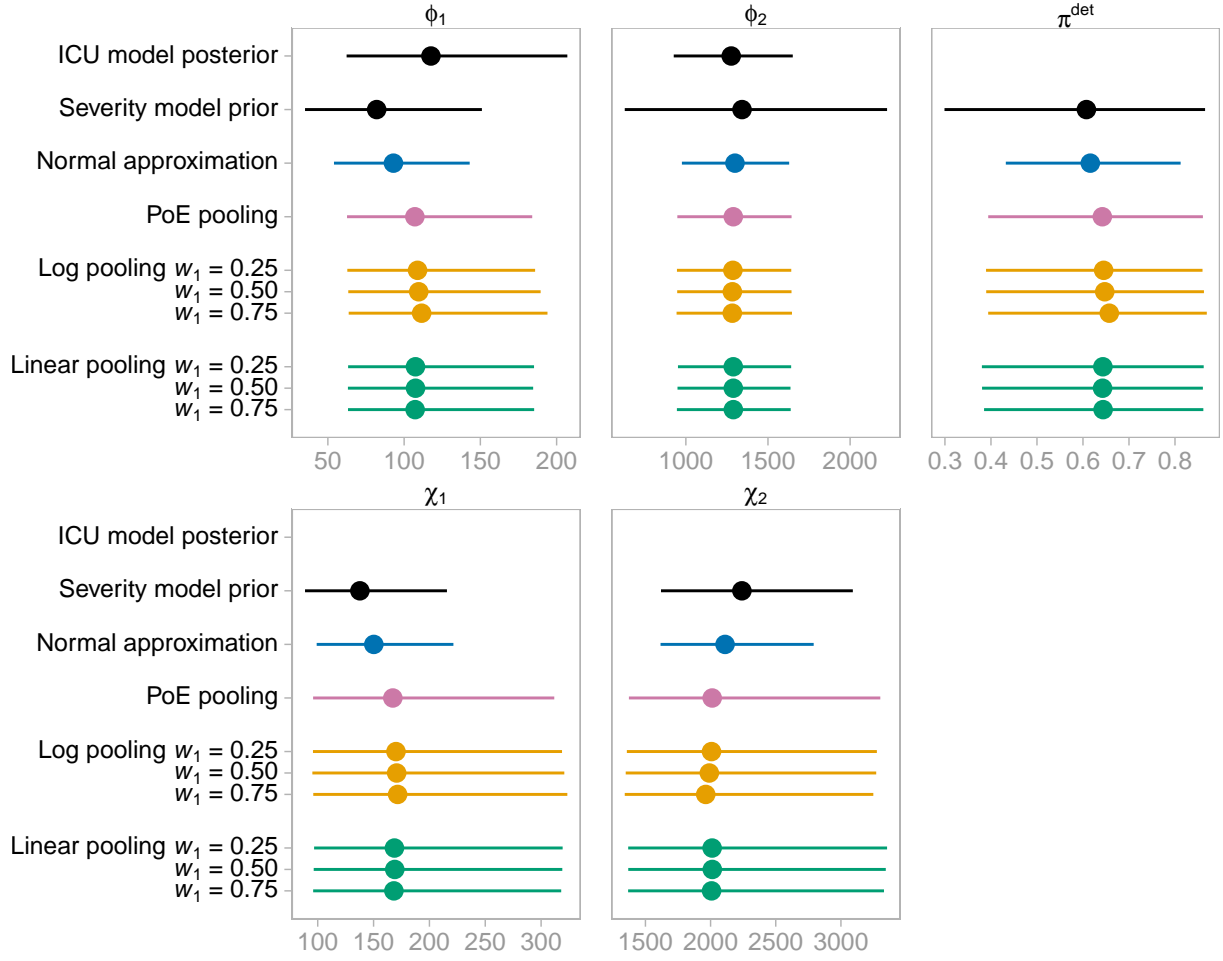


Figure 8: Medians and 95% credible intervals for: the posterior distribution for the link parameters ϕ_1 and ϕ_2 under the ICU submodel; the prior distribution for each parameter under the severity submodel; the posterior distribution for each parameter according to the normal approximation and the Markov melded model under each pooling rule.

The recovery rate is the probability that a bird that dies in year t is recovered. The probability of a bird released in year t_1 being never recovered is $\pi_{t_1,37} = 1 - \sum_{u=t_1+1}^{36} \pi_{t_1,u}$.

5.2.2 Census data

We assume that the observed census-type data x_t , which are available for 1965 ($t = 3$) to 1998, account for only breeding birds and that there is no emigration. We model the census data via the true number of breeding females $\mu_{A,t}$ and immature females $\mu_{C,t}$, and the productivity rate γ_t , the average number of female offspring per breeding female in year t . Specifically we assume for $t = 3, \dots, 36$

$$\begin{aligned} x_t &\sim \text{N}(\mu_{A,t}, \sigma^2) \\ \mu_{C,t} &\sim \text{Po}(\mu_{A,t-1}\gamma_{t-1}\eta_{C,t}) \\ \mu_{A,t} &\sim \text{Bin}(\mu_{C,t-1} + \mu_{A,t-1}, \eta_{A,t}), \end{aligned}$$

with the observation variance σ^2 assumed constant.

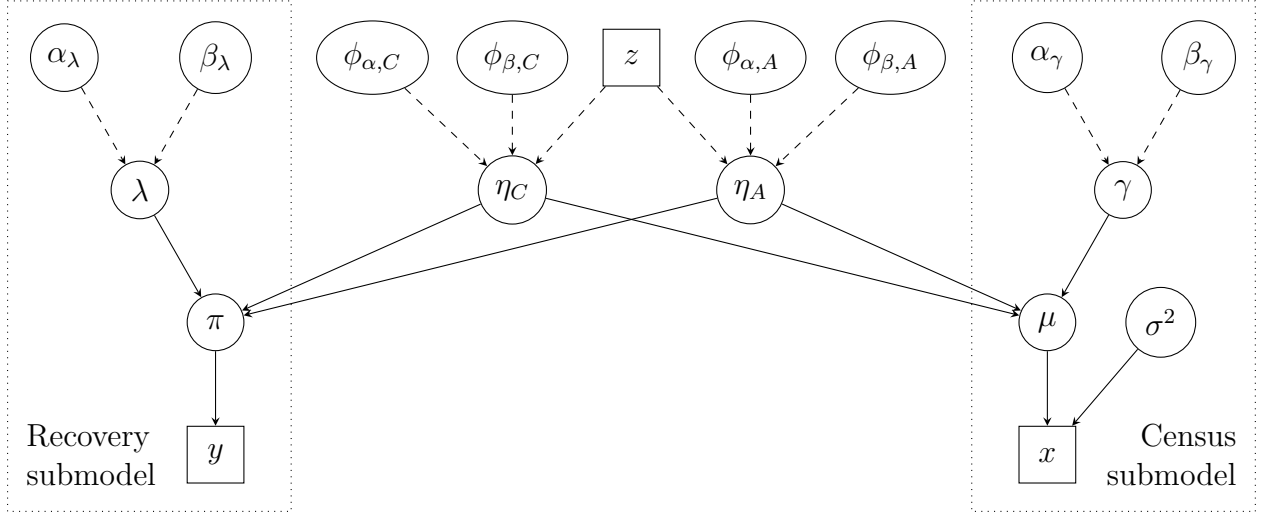


Figure 9: DAG representation of the joint ecology model. The recovery and census submodels are connected via the common parameter $\phi = (\phi_{\alpha,C}, \phi_{\alpha,A}, \phi_{\beta,C}, \phi_{\beta,A})$. For simplicity the time domain is suppressed: $y, x, \pi, \mu, \lambda, \eta_C, \eta_A, \gamma$ and z represent the collection of *all* quantities sharing the same variable name. For example, $\mu = \{\mu_{G,t} : G \in \{C, A\}, t = 3, \dots, 36\}$.

5.2.3 Regression models and prior distributions

We model the parameters $\eta_{G,t}$, λ_t and γ_t with regression models, with z_t denoting the (observed) number of frost days in year t .

$$\begin{aligned} \text{logit}(\eta_{G,t}) &= \phi_{\alpha,G} + \phi_{\beta,G} z_t & G = C, A \\ \text{logit}(\lambda_t) &= \alpha_\lambda + \beta_\lambda t \\ \log(\gamma_t) &= \alpha_\gamma + \beta_\gamma t \end{aligned}$$

We place lognormal priors on the number of immature females $\mu_{C,2}$ and breeding females $\mu_{A,2}$ in the year prior to our data series, with scale parameter 1 and location parameters $\mu_C = 200$ and $\mu_A = 1000$ respectively. We assume $\sigma^2 \sim \text{Inv-Gam}(0.001, 0.001)$ a priori, and independent $N(0, 10^2)$ prior distributions for all 8 regression parameters $(\phi_{\alpha,C}, \phi_{\alpha,A}, \alpha_\lambda, \alpha_\gamma, \phi_{\beta,C}, \phi_{\beta,A}, \beta_\lambda, \beta_\gamma)$.

5.2.4 Results

We split the joint model, as described in Section 3.5.2, into two components: the mark-recapture-recovery submodel and the census submodel. Denote by $\Omega_0 = (\eta_C, \eta_A, \phi_{\alpha,C}, \phi_{\alpha,A}, \phi_{\beta,C}, \phi_{\beta,A})$ the parameters shared by both submodels and by $\Omega_1 = (\pi, \lambda, \alpha_\lambda, \beta_\lambda)$ the parameters specific to the recovery submodel. Under both the mark-recapture-recovery submodel (stage one) and the census submodel (stage two), we use independent normal priors, with mean 0 and standard deviation $\sqrt{200}$, for each component $\phi_{\alpha,C}$, $\phi_{\alpha,A}$, $\phi_{\beta,C}$ and $\phi_{\beta,A}$ of the link parameter. These priors were chosen so that PoE pooling of these priors results in the original prior for the link parameters under the joint model.

In stage one we drew samples from the posterior distribution $p_1(\Omega_1, \Omega_0 \mid y)$ under the recovery submodel, and retained these samples for use as a proposal distribution in stage two, in which we drew samples under the full joint model. In stage one, we drew 2.5×10^5 MCMC iterations from the posterior distribution of the mark-recapture-recovery

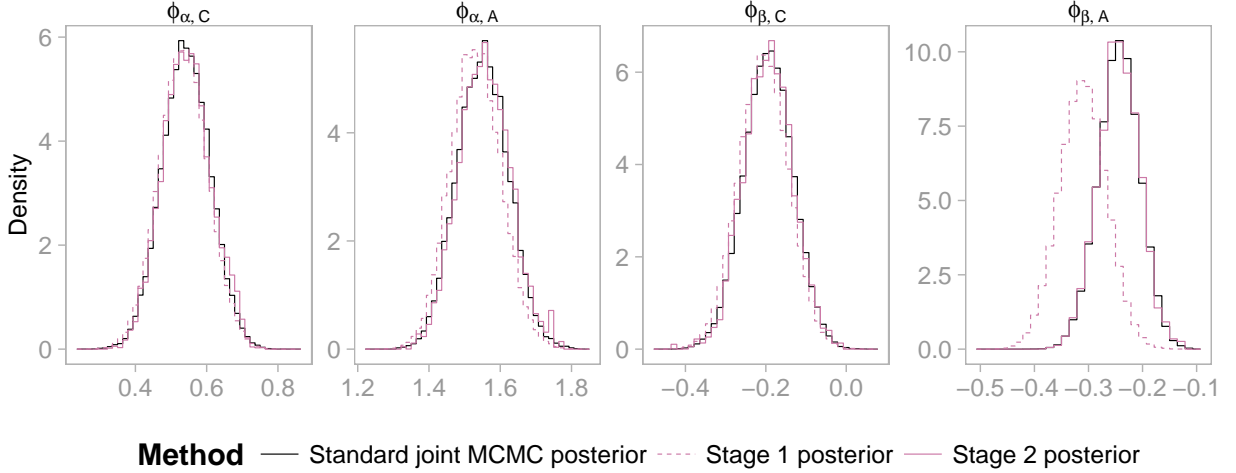


Figure 10: Histograms of the posterior densities of the link parameters $\phi_{\alpha,C}$, $\phi_{\alpha,A}$, $\phi_{\beta,C}$ and $\phi_{\beta,A}$ under the recovery submodel (Stage 1), and under the full joint model, as estimated by Stage 2 of the two stage sampler and by a standard MCMC sampler for the joint model.

submodel, taking 7 hours. In stage two, we discarded all but every 100th iteration, leaving 2.5×10^5 MCMC iterations for inference. This took $6\frac{1}{2}$ hours.

Figure 10 shows the results. We compare the two-stage estimates to estimates of the joint distribution based upon 6×10^5 MCMC iterations (retaining every 10th iteration) drawn using a standard (one stage) MCMC sampler, which took 22 hours to run in OpenBUGS. The components of the link parameter $\phi_{\alpha,C}$ and $\phi_{\beta,C}$ corresponding to the immature birds have posterior distributions that closely agree under the joint model and mark-recapture-recovery submodel alone, but there are differences in the parameters corresponding to mature birds. In particular there is a sizeable difference for the regression parameter $\phi_{\beta,A}$, which is estimated to be notably higher under the joint model than under the mark-recapture-recovery submodel alone. The two-stage approach accurately captures this shift (Figure 10, right-hand panel).

6 Further work and discussion

In this paper we present a unifying view and a generic method for joining and splitting probabilistic submodels that share a common variable. We demonstrate the practical advantages of our principled model surgery approach, with its key tools of marginal replacement and Markov melding, in the design of models and algorithms for applied data analyses. We have extended the notion of Markov combination to the case where marginal distributions in each submodel need not be identical. Note that the extent of similarity of the marginal distributions required to enable Markov melding is partly a subjective modelling judgement: the marginals should have at a minimum some level of overlap of support between the submodels. When the areas of support are far away from each other, such conflict should be assessed and resolved, for example through bias modelling (Turner *et al.*, 2009), before model joining is contemplated, for both interpretation and computational reasons (see below).

In this section we discuss connections to related literatures, alternative approaches, and potential future extensions of this approach.

6.1 Related work

Bayesian melding, decomposable graphical models. The key idea for a melding approach can be attributed to Poole and Raftery (2000). Their presentation, however, focuses on a limited set of models and is tied up with a deterministic link parameter ϕ . This tends to slightly obscure the key issues which we present more generally in the Markov melding framework of equation (2). Of course, the possibility of a deterministic dependency of ϕ needs to be taken into account, as we do in Supplementary Material A. We obtain results similar to Poole and Raftery (2000), but are able to clearly separate issues of marginal replacement from issues of a deterministic transformation of random variables. We also agree with their assessment that care is necessary in treating such transformations in order to avoid any form of the Borel paradox.

A further influence is the work on decomposable graphical models (Dawid and Lauritzen, 1993; Lauritzen, 1996), where a key concept is that of a separator, a subset of variables that splits the model into two parts that are independent conditional on the separator. The notion of a separator aligns with the concept of a link variable for Markov melding, where we are interested in the separation and conditional independence structure without insisting on any more specific independence structures as imposed by undirected graphical models. Furthermore, the rich literature on decomposable graphs and suitable algorithms, for example, junction tree algorithms (Lauritzen, 1996), suggests obvious extensions of Markov melding beyond the use of a single link variable to a series of link variables (separators) for joining several submodels into larger chain or tree formations. However, further research into the potential of such extensions and of practical implementation is required.

Marginal replacement. It is possible to rephrase a variety of modifications of Bayesian models as examples of marginal replacement (Section 3.3.1) emphasising its general applicability. For notational simplicity we drop normalising constants in what follows. For example, the cut operator (Lunn *et al.*, 2013b; Plummer, 2015a) replaces the joint marginal $p(\phi_1, \phi_2)$ of two variables linked by an edge $\phi_1 \rightarrow \phi_2$ in a directed graphical model by the marginal $p(\phi_1)$ (or more generally by $p(\phi_1)p(\phi_2 | \phi_1)^w$, $0 \leq w < 1$), thus cutting (or reducing) any *direct* flow of information from ϕ_2 back to ϕ_1 . The computational performance of an implementation of the cut operator via marginal replacement using empirical density estimation will be explored in future work.

Approximate Bayesian Computation (ABC) has been suggested in cases where standard Bayesian inference is difficult due to the computational complexity of evaluating the likelihood function (Tavaré *et al.*, 1997). As we show in Supplementary Material A, ABC can also be interpreted as a marginal replacement with an approximation of a point mass distribution as the new marginal.

Changing marginals by transformation. To join models with a link variable ϕ by Markov combination, the models need to agree in their marginal distributions on ϕ . We have proposed marginal replacement to achieve agreement (Section 3.3.1). An alternative approach is via a suitable transformation of ϕ so that all marginals agree. A similar technique is used when deriving multivariate distributions from copulas, which capture the dependency structure of a distribution but have uniform marginal distributions that can be transformed to any required target marginal (Durante and Sempi, 2010). Marginal replacement assumes that all models measure the link variable on the same scale but

differ in their prior distributions for this variable. In contrast, a transformation approach assumes the link variable is measured on a different scale in each model but there is agreement on its prior distribution between models: that is, after a suitable rescaling, all models agree in their marginal distributions on ϕ . This assumption can be used to set up suitable transformations (Supplementary Material D). The computational requirements for this alternative approach are similar to the ones presented for marginal replacement. In addition to kernel density estimation for the marginal distributions, estimates of cumulative distribution functions are required to define the transformations. However, there exists a rich literature on the latter problem (Durante and Sempi, 2010).

Approximate approaches Evidence synthesis models (Eddy *et al.*, 1992; Jackson *et al.*, 2009; Albert *et al.*, 2011; Commenges and Hejblum, 2012; Presanis *et al.*, 2014) often employ the approximate approach of summarising the results of a first-stage submodel via a Gaussian or other distribution, for use in a second-stage submodel as a likelihood term. We have demonstrated here that the approach is a particular case of our general Markov melding framework under a choice of PoE pooling, therefore justifying the approximation in joining submodels. A related field where an approximate approach has been widely used is in standard and network meta-analysis (for example, Hasselblad *et al.*, 1992; Ades and Sutton, 2006; Welton *et al.*, 2008): here effect size point estimates from each study enter a model through normal or other likelihood terms, with their corresponding estimated standard errors assumed fixed. Meta-analyses are, therefore, further examples where our Markov melding approach could be fruitfully applied.

Hierarchical models Similarly, in more general hierarchical models, splitting models to make inference faster or easier has previously been considered (Liang and Weiss, 2007; Tom *et al.*, 2010; Lunn *et al.*, 2013a). In this setting, posterior inference is first obtained from independent unit-specific submodels, with flat, independent priors replacing all hierarchical priors in the full, joint model. Inference for the full, joint model is recovered in stage two through Markov melding of these unit-specific submodels with dictatorial pooling, so that only the hierarchical prior is reflected in the final results. The approach makes it easier to consider more complex population models for the random effects, such as covariate selection models, and makes cross-validation across units more convenient (Goudie *et al.*, 2015). As well as Metropolis-Hastings samplers analogous to those considered in this paper, Liang and Weiss (2007) also proposed variants in which the stage one samples are resampled to enable a fully Gibbs sampler, and a variant in which samples in stage two are drawn according to a kernel density estimate of the stage one samples.

Tall data Our framework can also be viewed as encapsulating a range of approaches proposed in the ‘big data’ literature for computationally efficient methods when the number of observations is large (‘tall data’). Standard MCMC samplers are ill-suited to tall data because the acceptance probability, which needs to be evaluated at each MCMC iteration, typically depends on the entire data set. With tall data it may be infeasible even to store all of the data on a single computer, nevermind evaluate functions depending on the whole dataset thousands of times. A number of divide-and-conquer approaches in this setting have recently been proposed, in which the original exchangeable data y are partitioned into B batches y_1, \dots, y_B , each of which contains few enough observations that standard statistical methods can be applied without undue trouble. The key observation is that the full posterior distribution $p(\phi \mid y)$ can be split into a number of submodel

posteriors $p_b(\phi \mid y_b) \propto p(y_b \mid \phi)p(\phi)^{1/B}$, $b = 1, \dots, B$. This is equivalent to splitting the model in the manner we propose in this paper, with the original prior apportioned equally among the batches, and pooling with a PoE approach. Various approaches for integrating the batch-specific posteriors to approximate the overall posterior have been proposed (Huang and Gelman, 2005; Scott *et al.*, 2016; Neiswanger *et al.*, 2014; Wang and Dunson, 2013; Bardenet *et al.*, 2015; Minsker *et al.*, 2014), including approaches that are exact for normal posteriors, normal approximations, and non- and semi-parametric approaches to account for non-normality. However, this literature has so far only considered independent, identically distributed data, whereas here we have considered more general models and data.

Conflict assessment Splitting models into conditionally independent components at a set of separator or link parameters is also a key aspect of cross-validatory posterior predictive methods, including “node-splitting”, for assessing the consistency or otherwise of different subsets of evidence, whether prior or data (Presanis *et al.*, 2013; Gåsemysr and Natvig, 2009; Gåsemysr, 2016). As with any cross-validation technique (Goudie *et al.*, 2015), the systematic assessment of conflict at any node in a DAG (Presanis *et al.*, 2016) may be computationally challenging for larger models. Our Markov melding framework provides a natural modular approach for such a systematic model assessment, followed by fitting the full joint model, possibly after further model development to account for any detected conflict, while minimising the computational burden.

Weighting evidence Any prior pooling approach considered within our framework includes a judgement as to how to weight the different submodels. Various other methods have been proposed for weighting evidence, including the cut operator (Lunn *et al.*, 2013b; Plummer, 2015a); the power prior approach in clinical trials (for example, Neuenschwander *et al.*, 2009); and modularisation in the computer models literature (Liu *et al.*, 2009). We have already seen that the cut operator is an example of marginal replacement. However, further research is required to investigate the relationship of Markov melding to other weighting approaches. Note that in the power prior approach, attempts have been made to estimate the weight w given to each prior by treating w as an unknown parameter (Neuenschwander *et al.*, 2009), but the weight is only (strongly) identifiable if there is enough heterogeneity or conflict between the different submodels.

Supra-Bayesian approaches There are alternative methods to the pooling approach that we considered for combining priors. One is the supra-Bayesian approach (Winkler, 1968; Lindley *et al.*, 1979; Roback and Givens, 2001) in which the ultimate decision maker uses the elicited priors $p_m(\phi)$ to inform a model of expert opinion in the relevant domain. For example, Albert *et al.* (2012) elicited quantiles and probabilities for observable quantities from multiple experts, and combined them via a hierarchical model. Supra-Bayesian approaches require the ultimate decision maker to construct a plausible model of expert opinion. This could be challenging in the context we consider because the priors that we combine are the marginal distributions of the link variable in each submodel.

6.2 Computation

In the example of model splitting we present, we found significant speed-up in using the multi-stage sampling approach over a standard one-stage MCMC sampler. However, the

example is intended as an illustration of the methodology, and is not necessarily the most efficient approach for the particular ecology model we analysed. In particular, it is likely that faster implementations of the original joint model would be possible using custom compiled code, rather than generic BUGS-language software. The multi-stage sampling approach, however, is able to use the full range of submodel specifications permitted by the BUGS language. Such flexibility would be time-consuming to reproduce in custom software. The multi-stage sampling approach also opens the possibility of extending the submodels in ways that would lead to even a highly-optimised implementation of the full, joint model becoming too slow for practical use.

Density estimation. In the examples presented here the link variable is comparatively low dimensional and simple kernel density estimation using a heavy-tailed distribution such a multivariate t -distribution as kernel proved sufficient. Moreover, the results were very robust with respect to the choice of kernel parameters such as bandwidth. For higher-dimensional link variables more care in the choice of kernel estimation method might be required (Henderson and Parmeter, 2015). Alternatively, one might try to estimate the ratio of densities directly, which generally results in improved stability over ratio of density estimators (Sugiyama *et al.*, 2012). Some higher-level Bayesian modelling languages, such as the BUGS language or Stan (Carpenter *et al.*, In press), are comprehensive enough to support integration of empirical density estimates into the likelihood.

Multi-stage sampler. The multi-stage sampler broadly falls into the category of a sequential Monte Carlo (SMC) sampler (Doucet *et al.*, 2013). Some details on the connection are provided in Supplementary Material C. A key feature is that a distribution is carried forward to the next stage via samples drawn from it. However, there are also some significant differences to standard SMC schemes. On the one hand, while the submodel-specific parameters ψ_m could be equated to the latent state variables in sequential approaches, these state variables are typically linked by a transition distribution. Such direct link between parameters ψ_m is lacking here and there is still a need to sample these parameters outside the sequential scheme. The link parameter of interest, ϕ , on the other hand, is a static parameter and consequently difficult to estimate via sequential methods. Using a resample-move step (Gilks and Berzuini, 2001), which resamples parameters using transition kernels that leave the target distribution invariant, is less attractive in our case, since resampling from submodels from earlier stages would be involved. Our multi-stage approach is hence vulnerable to depletion of a sample representation of the distribution of ϕ in that fewer and fewer distinct values of ϕ are available at each stage. Schemes allowing an artificial evolution of the static parameter (for example, Liu and West, 2001) might be suitable for the multi-stage approach. However, unlike a typical sequential sampling scheme with a multitude of steps, our multi-stage approach for integrating different submodels is only intended for very few stages. In fact, all examples discussed here are two-stage examples, for which depletion is less likely to be an issue. Depletion of the sample also occurs when the posterior distributions of the joined submodels conflict.

An alternative but related approach to sampling a sequential model consisting of several components is taken in White *et al.* (2013). Posterior samples for the link variable are obtained from each component of the model and approximated by kernel densities. A product of these kernel densities serves as an approximation to the overall posterior distribution of the link variable. This approach could be adapted to our Markov melding setting as well.

References

- Ades, A. E. and Sutton, A. J. (2006) Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**, 5–35.
- Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K. and Rousseau, J. (2012) Combining expert opinions in prior elicitation. *Bayesian Analysis*, **7**, 503–532.
- Albert, I., Espié, E., de Valk, H. and Denis, J.-B. (2011) A Bayesian evidence synthesis for estimating Campylobacteriosis prevalence. *Risk Analysis*, **31**, 1141–1155.
- Bardenet, R., Doucet, A. and Holmes, C. (2015) On Markov chain Monte Carlo methods for tall data. arXiv:1505.02827.
- Besbeas, P., Freeman, S. N., Morgan, B. J. T. and Catchpole, E. A. (2002) Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. *Biometrics*, **58**, 540–547.
- Birrell, P. J., Ketsetzis, G., Gay, N. J., Cooper, B. S., Presanis, A. M., Harris, R. J., Charlett, A., Zhang, X.-S., White, P. J., Pebody, R. G. and De Angelis, D. (2011) Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *Proceedings of the National Academy of Sciences*, **108**, 18238–18243.
- Brooks, S. P., King, R. and Morgan, B. J. T. (2004) A Bayesian approach to combining animal abundance and demographic data. *Animal Biodiversity and Conservation*, **27**, 515–529.
- Buntine, W. L. (1994) Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, **2**, 159–225.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P. and Riddell, A. (In press) Stan: A probabilistic programming language. *Journal of Statistical Software*.
- Clark, J. S., Bell, D., Chu, C., Courbaud, B., Dietze, M., Hersh, M., HilleRisLambers, J., Ibáñez, I., LaDeau, S., McMahon, S., Metcalf, J., Mohan, J., Moran, E., Pangle, L., Pearson, S., Salk, C., Shen, Z., Valle, D. and Wyckoff, P. (2010) High-dimensional co-existence based on individual variation: a synthesis of evidence. *Ecological Monographs*, **80**, 569–608.
- Clemen, R. T. and Winkler, R. L. (1999) Combining probability distributions from experts in risk analysis. *Risk Analysis*, **19**, 187–203.
- Commenges, D. and Hejblum, B. P. (2012) Evidence synthesis through a degradation model applied to myocardial infarction. *Lifetime Data Analysis*, **19**, 1–18.
- Dawid, A. P. and Lauritzen, S. L. (1993) Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, **21**, 1272–1317.

- De Angelis, D., Presanis, A. M., Birrell, P. J., Tomba, G. S. and House, T. (2015) Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*, **10**, 83–87.
- Department of Health (2011) Department of Health Winter Watch. <http://winterwatch.dh.gov.uk>.
- Doucet, A., de Freitas, N. and Gordon, N., eds. (2013) *Sequential Monte Carlo Methods in Practice*. New York: Springer Science & Business Media.
- Draper, D. (1995) Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 45–97.
- Durante, F. and Sempi, C. (2010) Copula theory: an introduction. In *Copula Theory and its Applications* (eds. P. Jaworski, F. Durante, W. K. Härdle and T. Rychlik), pp. 3–31. Berlin: Springer-Verlag.
- Eddy, D. M., Hasselblad, V. and Shachter, R. (1992) *Meta-Analysis by the Confidence Profile Method*. London: Academic Press.
- Gåsemyr, J. and Natvig, B. (2009) Extensions of a conflict measure of inconsistencies in Bayesian hierarchical models. *Scandinavian Journal of Statistics*, **36**, 822–838.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014) *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC, 3rd edition.
- Gilks, W. R. and Berzuini, C. (2001) Following a moving target—Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 127–146.
- Goudie, R. J. B., Hovorka, R., Murphy, H. R. and Lunn, D. (2015) Rapid model exploration for complex hierarchical data: application to pharmacokinetics of insulin aspart. *Statistics in Medicine*, **34**, 3144–3158.
- Gåsemyr, J. (2016) Uniformity of node level conflict measures in Bayesian hierarchical models based on directed acyclic graphs. *Scandinavian Journal of Statistics*, **43**, 20–34.
- Green, P. J., Hjort, N. L. and Richardson, S. (2003) Introducing highly structured stochastic systems. In *Highly Structured Stochastic Systems* (eds. P. J. Green, N. L. Hjort and S. Richardson), pp. 1–12. Oxford: Oxford University Press.
- Hasselblad, V., Eddy, D. M. and Kotchmar, D. J. (1992) Synthesis of environmental evidence: nitrogen dioxide epidemiology studies. *Journal of the Air & Waste Management Association*, **42**, 662–671.
- Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
- Henderson, D. A., Boys, R. J. and Wilkinson, D. J. (2010) Bayesian calibration of a stochastic kinetic computer model using multiple data sources. *Biometrics*, **66**, 249–256.
- Henderson, D. J. and Parmeter, C. F. (2015) *Applied Nonparametric Econometrics*. New York: Cambridge University Press.

- Hinton, G. E. (2002) Training products of experts by minimizing contrastive divergence. *Neural computation*, **14**, 1771–1800.
- Huang, Z. and Gelman, A. (2005) Sampling for Bayesian computation with large datasets. Working paper.
- Jackson, C. H., Best, N. G. and Richardson, S. (2009) Bayesian graphical models for regression on multiple data sets with different variables. *Biostatistics*, **10**, 335–351.
- Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon Press.
- Liang, L.-J. and Weiss, R. E. (2007) A hierarchical semiparametric regression model for combining HIV-1 phylogenetic analyses using iterative reweighting algorithms. *Biometrics*, **63**, 733–741.
- Lindley, D. V., Tversky, A. and Brown, R. V. (1979) On the reconciliation of probability assessments. *Journal of the Royal Statistical Society: Series A (General)*, **142**, 146–180.
- Liu, F., Bayarri, M. J. and Berger, J. O. (2009) Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, **4**, 119–150.
- Liu, J. and West, M. (2001) *Combined parameter and state estimation in simulation-based filtering*, pp. 197–223. In Doucet *et al.* (2013).
- Lunn, D., Barrett, J., Sweeting, M. and Thompson, S. (2013a) Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**, 551–572.
- Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2013b) *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton: CRC Press.
- Massa, M. S. and Lauritzen, S. L. (2010) Combining statistical models. In *Contemporary Mathematics: Algebraic Methods in Statistics and Probability II* (eds. M. A. G. Viana and H. P. Wynn), pp. 239–260.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092.
- Miller, J. W. and Dunson, D. B. (2015) Robust Bayesian inference via coarsening. arXiv:1506.06101.
- Minsker, S., Srivastava, S., Lin, L. and Dunson, D. B. (2014) Robust and scalable Bayes via a median of subset posterior measures. arXiv:1403.2660v3.
- Müller, P. (1991) A generic approach to posterior integration and Gibbs sampling. Technical Report 91-09, Purdue University.
- Neal, R. (2003) Slice sampling. *Annals of Statistics*, **31**, 705–741.
- Neiswanger, W., Wang, C. and Xing, E. P. (2014) Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the Thirtieth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-14)*, pp. 623–632. Corvallis, Oregon: AUAI Press.

- Neuenschwander, B., Branson, M. and Spiegelhalter, D. J. (2009) A note on the power prior. *Statistics in Medicine*, **28**, 3562–3566.
- O’Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts’ Probabilities*. Chichester: John Wiley & Sons.
- Plummer, M. (2015a) Cuts in Bayesian graphical models. *Statistics and Computing*, **25**, 37–43.
- Plummer, M. (2015b) JAGS Version 4.0.1 user manual.
- Poole, D. and Raftery, A. E. (2000) Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association*, **95**, 1244–1255.
- Presanis, A. M., Ohlssen, D., Cui, K., Rosinska, M. and De Angelis, D. (2016) Conflict diagnostics for evidence synthesis in a multiple testing framework. Working paper.
- Presanis, A. M., Ohlssen, D., Spiegelhalter, D. J. and De Angelis, D. (2013) Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statistical Science*, **28**, 376–397.
- Presanis, A. M., Pebody, R. G., Birrell, P. J., Tom, B. D. M., Green, H. K., Durnall, H., Fleming, D. and De Angelis, D. (2014) Synthesising evidence to estimate pandemic (2009) A/H1N1 influenza severity in 2009–2011. *Annals of Applied Statistics*, **8**, 2378–2403.
- Public Health England (2014) Sources of UK flu data: influenza surveillance in the UK. <https://www.gov.uk/guidance/sources-of-uk-flu-data-influenza-surveillance-in-the-uk>.
- Roback, P. J. and Givens, G. H. (2001) Supra-Bayesian pooling of priors linked by a deterministic simulation model. *Communications in Statistics – Simulation and Computation*, **30**, 447–476.
- Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E. (2016) Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, **11**, 78–88.
- Sugiyama, M., Suzuki, T. and Kanamori, T. (2012) *Density Ratio Estimation in Machine Learning*. New York: Cambridge University Press.
- Tavaré, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.
- Tom, J. A., Sinsheimer, J. S. and Suchard, M. A. (2010) Reuse, recycle, reweigh: combating influenza through efficient sequential Bayesian computation for massive data. *Annals of Applied Statistics*, **4**, 1722–1748.
- Turner, R. M., Spiegelhalter, D. J., Smith, G. C. S. and Thompson, S. G. (2009) Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **172**, 21–47.

- Wang, X. and Dunson, D. B. (2013) Parallel MCMC via Weierstrass Sampler. arXiv:1312.4605.
- Welton, N. J., Cooper, N. J., Ades, A. E., Lu, G. and Sutton, A. J. (2008) Mixed treatment comparison with multiple outcomes reported inconsistently across trials: Evaluation of antivirals for treatment of influenza A and B. *Statistics in Medicine*, **27**, 5620–5639.
- Welton, N. J., Sutton, A. J., Cooper, N. J., Abrams, K. R. and Ades., A. (2012) *Evidence Synthesis for Decision Making in Healthcare*. Chichester: John Wiley & Sons.
- White, S. R., Kypraios, T. and Preston, S. P. (2013) Piecewise Approximate Bayesian Computation: fast inference for discretely observed Markov models using a factorised posterior distribution. *Statistics and Computing*, **25**, 289–301.
- Wilkinson, R. D. (2013) Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, **12**, 129–141.
- Winkler, R. L. (1968) The consensus of subjective probability distributions. *Management Science*, **15**, B61–B75.

Acknowledgements

This work was supported by the UK Medical Research Council [programme codes MC_UP_1302/3, MC_U105260556, U105260557 and MC_U105260799]. We are grateful to Ian White, Sylvia Richardson, Brian Tom, Michael Sweeting, Paul Kirk, Adrian Raftery, and the 2015 Armitage lecturers (Leonhard Held and Michael Höhle) for helpful discussions of this work. We also thank colleagues at Public Health England for providing data.

Supplementary Material

A Motivation for marginal replacement

We argue that a motivation for (3) is that p_{repl} minimises the Kullback-Leibler (KL) divergence of a distribution $q(\phi, \psi_m, Y_m)$ to $p(\phi, \psi_m, Y_m)$ under the constraint that the marginals on ϕ agree, $q(\phi) = p_{\text{new}}(\phi)$, that is,

$$p_{\text{repl}}(\phi, \psi_m, Y_m) = \operatorname{argmin}_q \{D_{\text{KL}}(q \parallel p_m) \mid q(\phi) = p_{\text{new}}(\phi) \text{ for all } \phi\}$$

This is easily shown as follows (we drop index m and variable Y for simplicity). The KL divergence under the constraint is given by

$$\begin{aligned} D_{\text{KL}}(q \parallel p) &= \int q(\phi, \psi) \log \frac{q(\phi, \psi)}{p(\phi, \psi)} d\phi d\psi \\ &= \int q(\psi \mid \phi) \log \frac{q(\psi \mid \phi)}{p(\psi \mid \phi)} d\psi q(\phi) d\phi + \int q(\phi, \psi) d\psi \log \frac{q(\phi)}{p(\phi)} d\phi \\ &= \int D_{\text{KL}}(q(\cdot \mid \phi) \parallel p(\cdot \mid \phi)) q(\phi) d\phi + \int p_{\text{new}}(\phi) \log \frac{p_{\text{new}}(\phi)}{p(\phi)} d\phi \end{aligned}$$

The second term is the KL divergence of the marginals and is constant. The first term can be minimised to 0 by choosing $q(\psi | \phi) = p(\psi | \phi)$ for all ϕ and consequently $q(\phi, \psi) = p(\psi | \phi)p_{\text{new}}(\phi)$ is the solution to the constrained KL divergence minimisation. A similar argument has been made in Poole and Raftery (2000) to justify their choice of a distribution for Bayesian melding. Notice that the same argument can still be made based on $D_{\text{KL}}(p_m \| q)$ with the roles of p_m and q exchanged.

Marginal replacement can also be seen as a generalisation of Bayesian updating in the light of new information. For example, if we learn that ϕ can assume only the value of a constant ϕ_0 , standard Bayesian updating entails conditioning on the new information ϕ_0 to form the posterior distribution $p_m(\psi_m, Y_m | \phi_0)$. This can be viewed as a special case of marginal replacement in which the new marginal $p_{\text{new}}(\phi)$ is the point mass $\delta_{\phi_0}(\phi)$ on ϕ_0 , since the marginal distribution of (ψ_m, Y_m) under the marginal replacement model is this posterior distribution, as follows by standard properties of the Dirac delta function δ_{ϕ_0} :

$$\begin{aligned} p_{\text{repl},m}(\psi_m, Y_m) &= \int p_{\text{repl},m}(\phi, \psi_m, Y_m) d\phi \\ &= \int p_m(\psi_m, Y_m | \phi) \delta_{\phi_0}(\phi) d\phi = p_m(\psi_m, Y_m | \phi_0) \end{aligned} \quad (9)$$

In this sense, (3) enables integration of new information on ϕ provided not only in the form of a specific value ϕ_0 but in the form of a general density function $p_{\text{new}}(\phi)$.

Finally, Approximate Bayesian Computation (ABC) can be interpreted as a marginal replacement similar to the replacement in equation (9), when ϕ typically represents a data variable¹. Instead of $p_{\text{new}}(\phi) = \delta_{\phi_0}(\phi)$ in standard posterior inference, ABC uses

$$p_{\text{new}}(\phi) = p(\phi) I(d(S(\phi), S(\phi_0)) < \epsilon)$$

where I is the indicator function of an event, d is a distance function, S some summary statistic for the data variable ϕ , ϕ_0 is observed and ϵ a small constant. ABC can thus be seen as very similar to standard posterior inference but with a widening of the δ function (Wilkinson, 2013; Miller and Dunson, 2015). In fact, the limits $\epsilon \rightarrow 0$ and $\epsilon \rightarrow \infty$ lead to the posterior and prior distributions on ϕ , respectively.

B Transformations with noninvertible deterministic functions

θ is a ℓ -dimensional real multivariate variable, and $\phi(\theta) = (\phi_1(\theta), \dots, \phi_k(\theta))$, $k < \ell$, a deterministic transformation that can be expanded to an invertible function $\phi_e(\theta) = (\phi(\theta), t(\theta))$, for $t(\theta) = (t_1(\theta), \dots, t_{\ell-k}(\theta))$. We assume the inverse mapping $\theta(\phi, t)$ and $\phi_e(\theta)$ have first derivatives. Mapping ϕ_e induces a probability distribution on (ϕ, t) which can be represented as

$$p(\phi, t) = p(\theta(\phi, t)) J_{\theta}(\phi, t) \quad (10)$$

where $J_{\theta}(\phi, t)$ is the Jacobian determinant for the transformation $\theta(\phi, t)$. The induced marginal distribution on ϕ can then be defined as

$$p(\phi) = \int p(\phi, t) dt \quad (11)$$

¹We thank Paul Kirk for this observation

Recall that the Jacobian determinant of the inverse transformation $\theta(\phi, t)$ is

$$J_\theta(\phi, t) = \left| \frac{\partial \theta / \partial \phi}{\partial \theta / \partial t} \right|_{(\phi, t)} = (J_{(\phi, t)}(\theta))^{-1} = \left| \frac{\partial(\phi, t)}{\partial \theta} \right|_{\theta(\phi, t)}^{-1} = \left| \frac{\partial \phi / \partial \theta}{\partial t / \partial \theta} \right|_{\theta(\phi, t)}^{-1}$$

where any $\partial u / \partial v = (\partial u_i / \partial v_j)_{ij}$ is the matrix of partial derivatives of functions u_i by variables v_j and $|\cdot|$ denotes the absolute value of the determinant.

Here we show that the value of $p(\phi)$ is independent of the particular parameterisation. That is, if $s(\theta)$ is an alternative parameterisation so that $\tilde{\phi}_e(\theta) = (\phi(\theta), s(\theta))$ also has an inverse mapping $\theta(\phi, s)$ then for a *fixed* ϕ we have an invertible transformation $s(t) = s(\theta(\phi, t))$ and

$$\begin{aligned} \int p(\phi, s) ds &= \int p(\phi, s(t)) \frac{ds}{dt}(t) dt = \int p(\theta(\phi, s(t))) \left| \frac{\partial \theta / \partial \phi}{\partial \theta / \partial s} \right|_{\theta(\phi, s(t))} \frac{ds}{dt}(t) dt \\ &= \int p(\theta(\phi, t)) \left| \frac{\partial \theta / \partial \phi}{\partial \theta / \partial s ds / dt} \right|_{\theta(\phi, t)} dt = \int p(\theta(\phi, t)) \left| \frac{\partial \theta / \partial \phi}{\partial \theta / \partial t} \right|_{\theta(\phi, t)} dt \\ &= \int p(\phi, t) dt = p(\phi) \end{aligned}$$

where we used the multilinearity of the determinant, the chain rule for multidimensional derivatives, and that $\theta(\phi, s(t)) = \theta(\phi, t)$ by the definition of $s(t)$. Consequently, $p(\phi)$ as the induced probability on all values taken by $\phi(\theta)$ is well defined.

The marginal distribution (11) of ϕ in the density $p(\phi, t)$ in (10) can then be replaced by any other desired marginal $p_{\text{new}}(\phi)$ via marginal replacement as in (3)

$$p_{\text{repl}}(\phi, t) = \frac{p(\phi, t)}{p(\phi)} p_{\text{new}}(\phi) = \frac{p(\theta(\phi, t)) J_\theta(\phi, t)}{p(\phi)} p_{\text{new}}(\phi)$$

Finally, the new distribution $p_{\text{repl}}(\phi, t)$ is mapped back to θ using the invertible mapping $\phi_e(\theta) = (\phi(\theta), t(\theta))$

$$p_{\text{repl}}(\theta) = p_{\text{repl}}(\phi(\theta), t(\theta)) J_\theta(\phi, t)^{-1} = p(\theta) \frac{p_{\text{new}}(\phi(\theta))}{p(\phi(\theta))} \quad (12)$$

which results in equation (4).

The last equation is similar to equation (16) in Poole and Raftery (2000). One of the key issues in their study is how to distribute the probability density at $p_{\text{new}}(\phi_0)$ over θ with $\phi(\theta) = \phi_0$. Equation (16) in Poole and Raftery (2000) as well as equation (12) here suggest doing this in proportion $p(\theta)/p(\phi(\theta))$ of the contribution of density $p(\theta)$ to $p(\phi(\theta)) = p(\phi_0)$. Poole and Raftery (2000) justify this approach more directly by using Kullback-Leibler divergence similar to Supplementary Material A above. Here it is a consequence of our slightly more general marginal replacement framework, which can also be justified by a Kullback-Leibler divergence argument as in Supplementary Material A.

C Multi-stage and sequential Monte Carlo sampling

A high-level feature of a sequential Monte Carlo (SMC) approach (Doucet *et al.*, 2013) is the aim to obtain a sample S_M from a distribution π_M via sampling from intermediate distributions π_1, \dots, π_M producing samples S_1, \dots, S_M , where $S_{\ell-1}$ is used to produce S_ℓ .

In this broad sense our multi-stage sampler is an example of such an algorithm. There are, however, deviations from a typical implementation of an SMC approach.

Formally, our target distributions are

$$\pi_\ell(\phi, \psi_1, \dots, \psi_\ell) \propto p_{\text{meld},\ell}(\phi, \psi_1, \dots, \psi_{\ell-1}, y_1, \dots, y_{\ell-1}) = \prod_{m=1}^{\ell} \rho_m(\phi, \psi_m)$$

as in (6). For simplicity we assume we are mostly interested in tracking samples of ϕ through the stages: $S_\ell = \{\phi_\ell^{(1)}, \dots, \phi_\ell^{(n_M)}\}$. If parameters ψ_m can be marginalised over, one could employ a typical sequential importance sampling scheme: sample S_ℓ from $S_{\ell-1}$ with probability proportional to

$$w_\ell^{(i)} = \frac{\pi_\ell(\phi_{\ell-1}^{(i)})}{\pi_{\ell-1}(\phi_{\ell-1}^{(i)})} = \rho_\ell(\phi_{\ell-1}^{(i)})$$

Note that we only need to evaluate the likelihood $\rho_m(\phi)$ for the last submodel $m = \ell$ due to the factorisation of p_{meld} . Equivalently, a sample can be obtained via Metropolis-Hastings sampling with target-to-proposal density ratio

$$R(\phi^\star, \phi) = \pi_\ell(\phi^\star) \times \frac{1}{q(\phi^\star)} = \pi_\ell(\phi^\star) \times \frac{1}{\pi_{\ell-1}(\phi^\star)} = \rho_\ell(\phi^\star)$$

where the proposal functions $q(\phi^\star)$ just samples uniformly from $S_{\ell-1}$.

We opted for the latter sampling approach since for the models envisaged it is rarely possible to marginalise out ψ_m and the Metropolis-Hastings sampler is able to sample from both ϕ and ψ_m together.

A notorious problem with static parameters such as ϕ is depletion of the sample with fewer distinct values of ϕ at each stage. Various schemes have been proposed to rejuvenate the sample. Liu and West (2001) propose adding a disturbance ζ to ϕ at each stage. This amounts to sampling from a kernel smoothed version of the original sample. Care needs to be taken to avoid undue increase in variance from stage to stage. Liu and West (2001) show how the increase can be controlled by cleverly correlating the disturbance ζ with ϕ .

Gilks and Berzuini (2001) propose a rejuvenation through a move step after the sampling step. This move step, applied at stage ℓ , needs to leave distribution π_ℓ invariant, for example, by one or more Metropolis-Hastings steps. In our case this is only possible by evaluating the full distribution $p_{\text{meld},\ell}$ involving all submodels $m = 1, \dots, \ell$, somehow defeating the purpose of the scheme to avoid revisiting submodels earlier than ℓ . However, for a long sequence of submodels an occasional move step might be beneficial despite the increase in computation.

D Transformation of marginals of the link variable

An alternative approach to achieve the same distribution of the link variable ϕ in all models required by (1) is via a suitable transformation of ϕ so that all marginals agree similar to a copula approach (Durante and Sempì, 2010). We assume we have link variables ϕ_m for each model m , measuring the same quantity (for example weight) but on different scales (say, kilograms, stones, pounds), which we indicate by a model-specific index m of ϕ_m . However, the twist is that we cannot assume the transformations between scales are known. Instead we assume they can be reconstructed by matching quantiles of the prior

distributions on the link variables. That is, find transformations so that all distributions are identical after rescaling.

We further assume we have a presentation of the link variable ϕ on a standard scale with a target distribution $p(\phi)$. For a suitable transformation for model m , let $F_m(\phi_m)$ and $F_{\text{new}}(\phi)$ denote the cumulative distribution function for $p_m(\phi_m)$ and $p_{\text{new}}(\phi)$ and let $F_m^-(\phi)$ and $F_{\text{new}}^-(\phi)$ denote their inverse functions. The model-specific mappings $\phi_m = Q_{m,\text{new}}(\phi) = F_m^-(F_{\text{new}}(\phi))$ transform between ϕ_m and ϕ preserving their distributions $p_m(\phi_m)$ and $p(\phi)$:

$$p_m(\phi_m) = p_m(Q_{m,\text{new}}(\phi)) \frac{dQ_{m,\text{new}}}{d\phi}(\phi) = p_m(Q_{m,\text{new}}(\phi)) \frac{p_{\text{new}}(\phi)}{p_m(Q_{m,\text{new}}(\phi))} = p_{\text{new}}(\phi) \quad (13)$$

We are now able to define new distributions that agree in their marginals on ϕ_m and ϕ by applying transformations $\phi_m = Q_{m,\text{new}}(\phi)$ to $p_m(\phi_m, \psi_m, Y_m)$

$$\begin{aligned} p_{\text{trans},m}(\phi, \psi_m, Y_m) &= p_m(Q_{m,\text{new}}(\phi), \psi_m, Y_m) \frac{dQ_{m,\text{new}}}{d\phi}(\phi) \\ &= \frac{p_m(Q_{m,\text{new}}(\phi), \psi_m, Y_m)}{p_m(Q_{m,\text{new}}(\phi))} p_{\text{new}}(\phi) \\ &= p_m(\psi_m, Y_m \mid Q_{m,\text{new}}(\phi)) p_{\text{new}}(\phi) \end{aligned} \quad (14)$$

Applying Markov combination to these transformed models results in a joint distribution

$$\begin{aligned} p_{\text{meldtrans}}(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M) &= p_{\text{new}}(\phi) \prod_{m=1}^M p_m(\psi_m, Y_m \mid Q_{m,\text{new}}(\phi)) \\ &= p_{\text{new}}(\phi) \prod_{m=1}^M \frac{p_m(Q_{m,\text{new}}(\phi), \psi_m, Y_m)}{p_m(Q_{m,\text{new}}(\phi))} \end{aligned} \quad (15)$$

Remarkably, it is straightforward to show that the choice of distribution $p_{\text{new}}(\phi)$ has no influence on $p_{\text{meldtrans}}$, it is only a convenient way to define the required transformations. If $p_{\text{alt}}(\phi_{\text{alt}})$ is an alternative target distribution with cumulative distribution function F_{alt} we define the transformation $\phi = Q_{\text{new,alt}}(\phi_{\text{alt}}) = F_{\text{new}}^-(F_{\text{alt}}(\phi_{\text{alt}}))$ which preserves marginals on ϕ and ϕ_{alt} . When we define $Q_{m,\text{alt}}(\phi_{\text{alt}}) = F_m^-(F_{\text{alt}}(\phi_{\text{alt}}))$ we also have $Q_{m,\text{alt}}(\phi_{\text{alt}}) = Q_{m,\text{new}}(Q_{\text{new,alt}}(\phi_{\text{alt}}))$ and so

$$p_m(\psi_m, Y_m \mid Q_{m,\text{new}}(\phi)) = p_m(\psi_m, Y_m \mid Q_{m,\text{alt}}(\phi_{\text{alt}}))$$

Similar to (13) for a transformation of ϕ_{alt} to $\phi = Q_{\text{new,alt}}(\phi_{\text{alt}})$ we have $p_{\text{alt}}(\phi_{\text{alt}}) = p_{\text{new}}(\phi)$ and (15) becomes

$$p_{\text{meldtrans}}(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M) = p_{\text{alt}}(\phi_{\text{alt}}) \prod_{m=1}^M p_m(\psi_m, Y_m \mid Q_{m,\text{alt}}(\phi_{\text{alt}}))$$

The influence of other submodels on submodel m in the joint model $p_{\text{meldtrans}}$ can now be made explicit easily by setting $p_{\text{alt}} = p_m$

$$\begin{aligned} p_{\text{meldtrans}}(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M) &= p_m(\phi_m) p_m(\psi_m, Y_m \mid \phi_m) \prod_{\ell \neq m} p_\ell(\psi_\ell, Y_\ell \mid Q_{\ell,m}(\phi_m)) \\ &= p_m(\phi_m, \psi_m, Y_m) \prod_{\ell \neq m} p_\ell(\psi_\ell, Y_\ell \mid Q_{\ell,m}(\phi_m)) \end{aligned} \quad (16)$$

with $\phi_\ell = Q_{\ell,m}(\phi_m) = F_\ell^-(F_m(\phi_m))$ a transformation that preserves the marginals $p_m(\phi_m)$ and $p_\ell(\phi_\ell)$.

Equation (16) shows that the distribution of ϕ_m in submodel m is influenced only through the likelihoods of the transformed variable $\phi_\ell = Q_{\ell,m}(\phi_m)$ in the other submodels. The transformations relate the ϕ_ℓ so that the quantiles of the distributions $p_\ell(\phi_\ell)$ of all models ℓ match.

This form of melding is useful when it is assumed that the marginals of the joint variable are specified correctly for each submodel and are essentially the same, but the variable is expressed on a different scale in each model and transformations to a common scale are needed to reveal the common underlying marginal. Also observe that, although changing p_{new} has no influence on $p_{\text{meldtrans}}$, for computational reasons it should be chosen so that transformations $Q_{m,\text{new}}$ can be estimated easily.